Statistics for Data Analysis - TABA

Saranya Varshni Roshan Karthikha
StudentId: 20154801
MSc in Data Analytics
National College of Ireland Dublin, IRELAND
Email: x20154801@student.ncirl.ie. URL: www.ncirl.ie

***Abstract*—This technical write-up focuses on experimenting the statistical analysis of - Time series and Logistic Regression. Prediction of future data points plays a vital role in every domain. Be it a pandemic or a stock market historical data set, forecasting the future range or value keeps us better prepared for the near future and also helps in making critical business decisions. This project is motivated to deal with two time series data sets that is overseas trip data set , new house registration and one logistic regression with child birth data set.**

Our aim is to use six different time series modeling for forecasting the future data set namely, Random walk, Holt's, Naive Bayes.

In Logistic regression, Principal Component Analysis is used as a technique to reduce the dimensions and further analyse on the regression.

***Index Terms*—Time series, trend, forecast, Random walk, Holt's, Naive, Holt's winter**

I. Introduction

Forecasting the demand in future is significant because this will make us to plan efficiently the resources for future requirements. The observations are recorded in a chronological order. These records are spaced with in a time span. when plotted they exhibit a pattern or trend in a specific direction or they are random.

Time series analysis and Logistic regression were implemented using R language and IBM SPSS respectively.

II. Time Series

Statistical methodology utilized to analyse series of data to identify general cyclic behaviours, patterns, trends, whether the data is seasonal or non-seasonal. These observations are noted with respect to time. The primary goal of this analysis is to identify characteristics which can be used in deriving meaningful statements and outcomes in future. Also prominently auto correlation, partial auto correlation and line graphs are used to depict and showcase the future events.

Fundamentally there are three different methodologies in time series.

1) Moving Average It is a simple and basic method that forecasts the next data point to be the mean of all previous observations. This can also help in visualizing the trends before decomposing the data set.

2) Exponential Smoothing Weight of each observation is decreased in accordance to the smoothing factor. This smooths the extreme peaks and trough in our data series. This also corresponds to double and triple exponential with the increase of smoothing factor value.

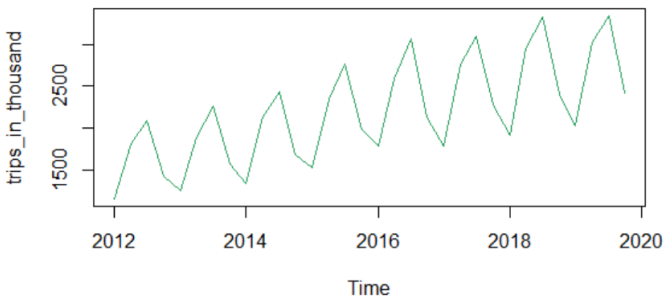

Fig. 1. Plot for Overseas Trip data

3) ARIMA Autoregressive moving average can be briefed as the combination of simple models to make up to a complex forecast model. This finds a relationship between observations and lagged observations

There are four main components in time series :

1)Trend: It represents the general trend of data over a long period of time. For instance if there was a steady increase in the stock market price, then an upward trend can be observed overall even though there are n number of peaks and troughs in the graph.

2)Cyclic: This behaviour represents the cyclic nature of the graph that occurs more than a year and this is different from seasonal pattern

3)Seasonal effect: Seasonal pattern represents the repetitive behaviour of the data points within a year.

4)Irregular Fluctuation: The uncertainty in the data that is the irregular and unexplained fluctuation.

A. Overseas Trip

- **Data Description:** This data set contains details of overseas trips made to Ireland by the non residents. Frequency of the data is observed to be quarterly. And they range from 2012 to 2019 with trips made in units thousand.
- **Assessing the components of Time series:** Figure 1 presents a line plot that depicts the overall pattern as upwards trend and high seasonality can be identified in the overseas data provided. The seasonality remains the same for all years but the trend is moving in upward direction.

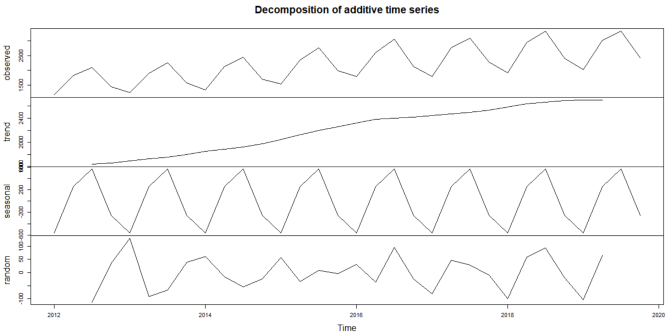

Fig. 2. Decomposed Overseas Trip data

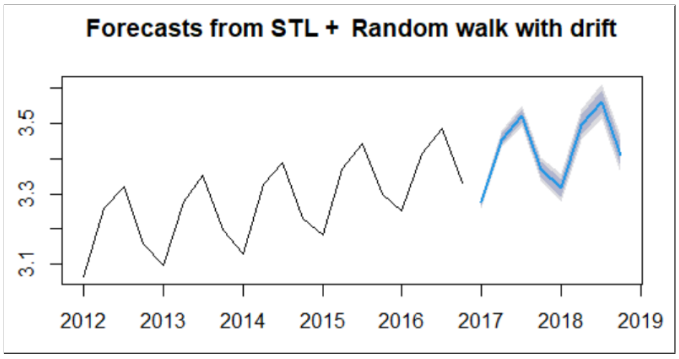

Fig. 3. Forecast using Random Walk

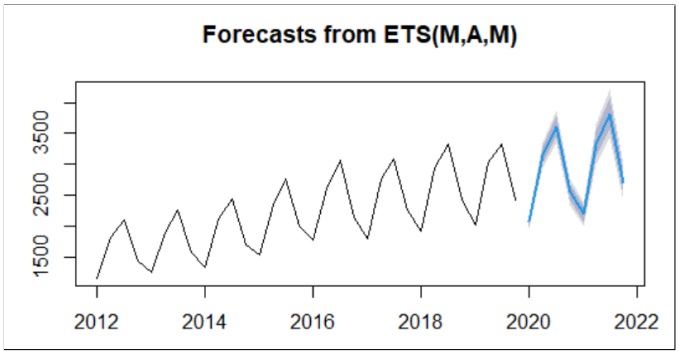

Fig. 4. Forecast using Random Walk

Time series is the combination of different components like trend, seasonality, cyclic behaviour along with the irregularity in data.

Decomposing the time series helps us understand the different components in a distinct way. It also improves the forecast accuracy.

From figure 2 it can be inferred that after decomposition the trend is upwards and a regular seasonality for every year.

The progress in data can be of two types :

1) Additive: This refers to the plot with constant mean and variance throughout the time span.

2) Multiplicative: Multiplicative represents variations in mean of the value throughout the series time line. The mean and variance might be progressively increasing towards the end of time series.

- **Time Series Models:**

Once the data set was imported into a list of values, to perform time series analysis, we need to convert it into a time series object. This was done using ts() where the start period and frequency was specified as parameters. Descriptive statistics was performed to understand more on the data.

Some exploratory analysis that was made at this point in time :

1)Autocorrelation: We checked which value of past observation matches with the current value.

2)Spectral analysis: Performed for cyclic behaviour observation in the data set provided.

3)Estimating Trend and Decomposition:

Decomposition of data is done to view all the components in different categories.

4)Achieving Stationary:

Stationary behaviour is necessary in a data set before performing analysis of time series on the same.

– Difference was computed with lag of 1 to remove the trend

– log function was applied to stabilize the variance

– Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test was performed to find out the difference or the regression on the data to make it stationary.

A. Random Walk:

Random walk model specifies that the distribution in the data series are independent of each other and it is very unlikely to predict the future values with the analysis of historical data. Figure 4 represents the Forecast made using random walk model.

B. Holt's Model:

This method belongs to the exponential smoothing methodologies. generally it is applied for series with linear trend. Three factors in this Holt's model are : Alpha - to weight the past and recent observation Beta - smoothing of the trend Gamma - coefficient for the seasonal smoothing Figure 4 represents forecasting using Holt's model.

C. Seasonal Naive:

This model believes that the forecast can be made with the last value of the same season of the year will be repeated. Figure 5 represents forecasting using Seasonal Naive model.

D. ARIMA :

We have used ARIMA model for our data set. There are totally three subdivisions in ARIMA model where AR stands for Autoregression and MA stands for Moving average while I stands for Integration. On execution of this AR we get the P value and from Moving average we get the q value and D is obtained on difference measure. If the model is seasonal then use nsdiffs() and if not

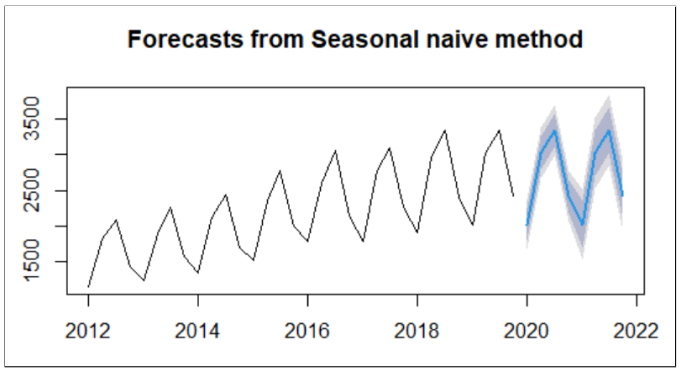

Fig. 5. Forecast using Seasonal Naive

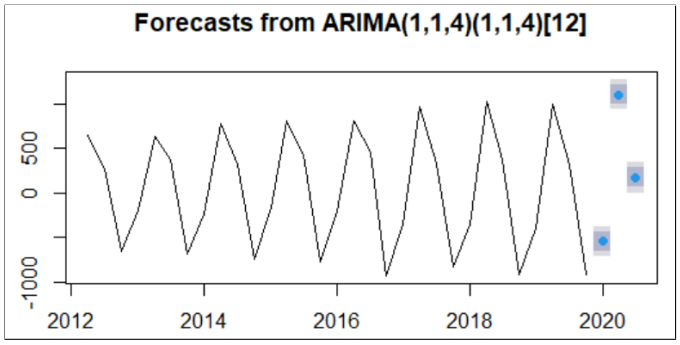

Fig. 6. Forecast using ARIMA

seasonal then use ndiffs() to identify the same.

- **Diagnostic Measures:**
  P -value in Dickey fuller test can be used to determine whether the time series is stationary or not. It ranges from 0 to 1. Figure 9 represents the p value as 0.99 before making the data stationary and figure 10 shows p value

```
> forecast(fitARIMA,h=3)
        Point Forecast      Lo 80     Hi 80       Lo 95     Hi 95
2020 Q1      -540.4033 -648.74066 -432.0659 -706.091008 -374.7155
2020 Q2      1099.9635  990.05865 1209.8684  931.878533 1268.0485
2020 Q3       177.4608   63.44583  291.4757    3.089964  351.8316
> plot(forecast(fitARIMA,h=3))
> coeftest(fitARIMA)

z test of coefficients:

       Estimate Std. Error  z value  Pr(>|z|)
ar1  -0.9399292  0.0920052 -10.2160 < 2.2e-16 ***
ma1   0.0074652  0.4285017   0.0174 0.9861003
ma2  -1.2863813  0.3463673  -3.7139 0.0002041 ***
ma3   0.0089345  0.4259171   0.0210 0.9832640
ma4   0.9990970  0.4586973   2.1781 0.0293972 *
sar1  0.2144347  1.0442498   0.2053 0.8373002
sma1  1.1653592  3.7445976   0.3112 0.7556404
sma2 -0.0180609  3.2641106  -0.0055 0.9955852
sma3 -1.1759755  3.1849246  -0.3692 0.7119549
sma4 -0.9654028  3.7710212  -0.2560 0.7979465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> accuracy(fitARIMA)
                     ME     RMSE      MAE       MPE     MAPE       MASE
Training set -2.970025 30.75043 18.25362 1.726677 3.147984 0.02215425
                    ACF1
Training set -0.1655082
```

Fig. 7. Model accuracy and forecasting

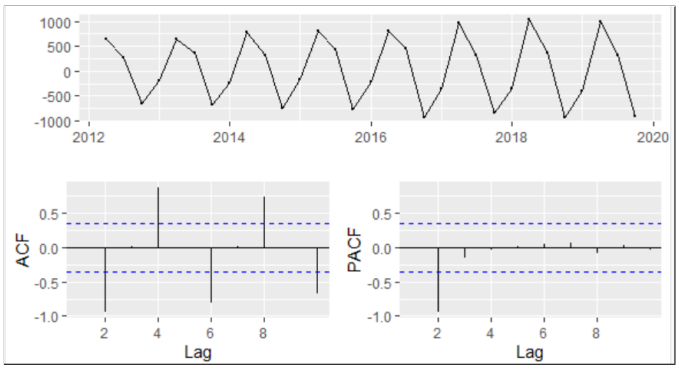

Fig. 8. Autocorrelation and Partial Autocorrelation for determining P and Q values

```
        Augmented Dickey-Fuller Test

data:  dt
Dickey-Fuller = 0.023128, Lag order = 3, p-value = 0.99
alternative hypothesis: stationary

Warning message:
In adf.test(dt) : p-value greater than printed p-value
```

Fig. 9. Dickey Fuller test before making the data set stationary

got reduced to 0.4 after the reduction of non - stationary elements.

- **Forecast Made:**
  Figures 11, 12 and 13 represents the summary from Holt's, Seasonal Naive, and ARIMA models implemented. RMSE of 54.698, 176.650 and 30.750 in Holt's, Seasonal Naive, and ARIMA models. Among which ARIMA stands to be the best one.

B. New House Registration

- Data Description:
  It is a series of new house registration made annually from year 1978 till 2019. Total observations count is 42.
- Assessing the components of Time series:
  With reference to the figure 14 it can be observed that there is no seasonality or general trend. Also with the given time span there is no cyclic behaviour of data identified.
- Time Series Model
  1) SES (Simple Exponential Smoothing):
  It is a forecasting methodology implemented for univariate attribute that has no trend and no seasonality, which is a perfect match for this data set.

```
        Augmented Dickey-Fuller Test

data:  dt_diif
Dickey-Fuller = -2.4204, Lag order = 3, p-value = 0.4109
alternative hypothesis: stationary
```

Fig. 10. Dickey Fuller test after making the data set stationary

```
ETS(M,A,M)

Call:
 ets(y = dt, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.6696
    beta  = 0.0119
    gamma = 1e-04

  Initial states:
    l = 1534.679
    b = 42.6526
    s = 0.8794 1.2604 1.1155 0.7447

  sigma:  0.0269

     AIC     AICc      BIC
378.8267 387.0085 392.0183

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE
Training set -7.22914 54.69822 44.54334 -0.3013343 2.017642 0.2906987
                    ACF1
Training set -0.05187753
```

Fig. 11. Summary for Holt's method

```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = dt, h = 8)

Residual sd: 176.6505

Error measures:
                    ME     RMSE      MAE      MPE     MAPE MASE      ACF1
Training set 153.2286 176.6505 153.2286 6.975085 6.975085    1 0.5355186
```

Fig. 12. Summary for Seasonal Naive method

```
> summary(fitARIMA)

Call:
arima(x = dt_diif, order = c(1, 1, 4), seasonal = list(order = c(1, 1, 4), perio
d = 12),
    method = "ML")

Coefficients:
          ar1     ma1      ma2     ma3     ma4    sar1    sma1     sma2
      -0.9399  0.0075  -1.2864  0.0089  0.9991  0.2144  1.1654  -0.0181
s.e.   0.0920  0.4285   0.3464  0.4259  0.4587  1.0442  3.7446   3.2641
         sma3     sma4
      -1.1760  -0.9654
s.e.   3.1849   3.7710

sigma^2 estimated as 1628:  log likelihood = -110.79,  aic = 243.59

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE       MASE
Training set -2.970025 30.75043 18.25362 1.726677 3.147984 0.02215425
                    ACF1
Training set -0.1655082
```

Fig. 13. Summary for ARIMA Model

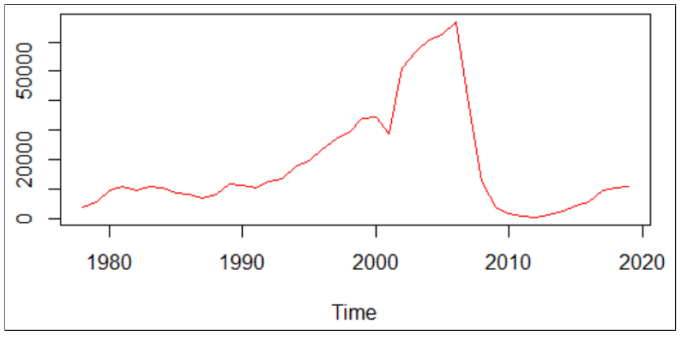

Fig. 14. Plot for new house registration Model

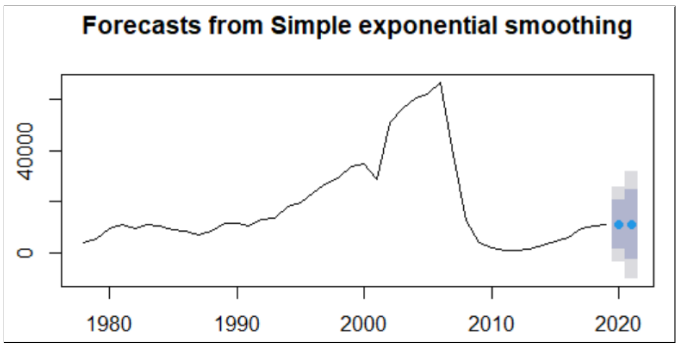

Fig. 15. Plot for SES model

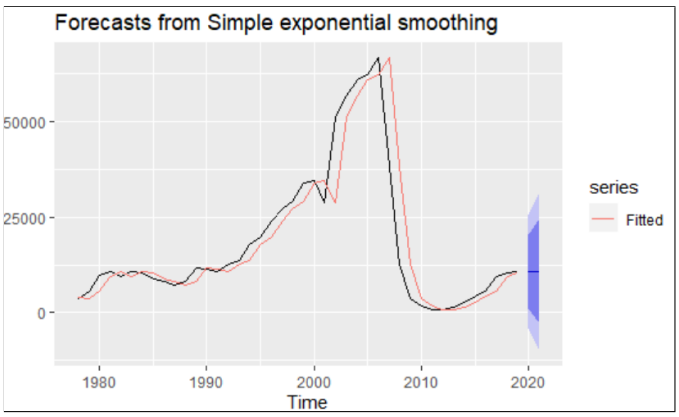

Fig. 16. Plot for SES fit Model

Figure 15 represents the forecasting that was made with SES and followed by the model fitting graph in figure 16

Accuracy for the same is printed in the figure 17 where the MAPE is 33.33.

2) Winter Holt's: This model helps in diagnosing the non-linear model. Figure 19 represents the forecasting that was made using with the damp and without the damp fit line.

3) ARIMA:

Another step in the analysis is the Ljung-Box test. The H0 of the Ljung-Box test specifies that the residuals are independently distributed. The p-value of the Ljung Box test we obtained proves that we can accept null hypothesis and this model is a good fit.

- Forecast Made:

  Figure 24, 25, 26 represents the summary for the chosen 3 models.

```
> round(accuracy(ses_new_house_reg),2)
                     ME     RMSE      MAE   MPE  MAPE MASE ACF1
Training set 165.21 7378.82 3875.79 -7.77 33.33 0.98 0.42
```

Fig. 17. Plot for SES accuracy

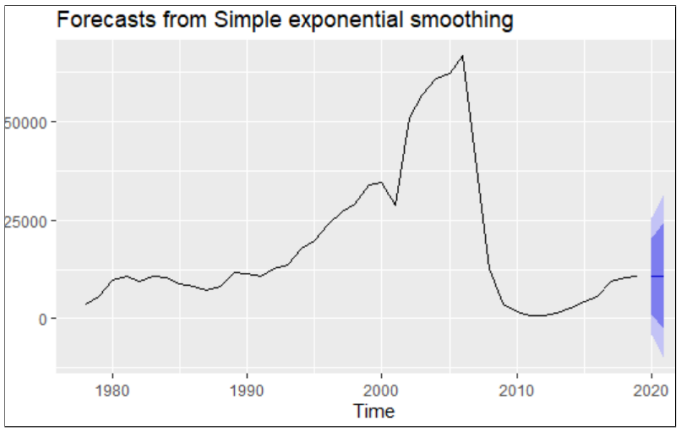

Fig. 18. Plot for SES forecast

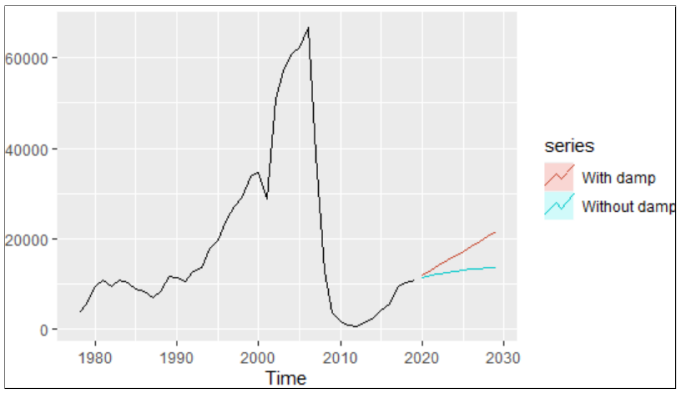

Fig. 19. Plot for HOLT's with and without damp

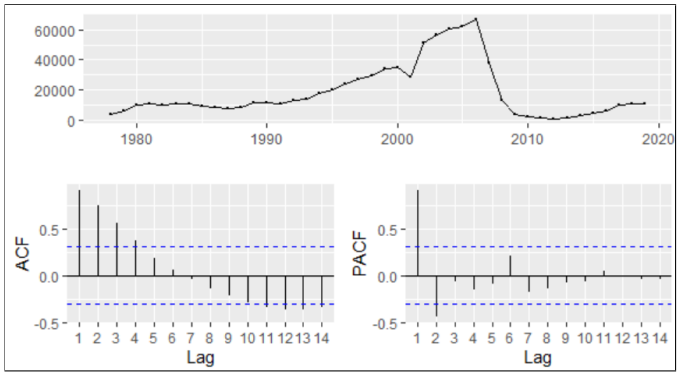

Fig. 20. Plot for ARIMA

```
> Box.test(fit$residuals, type = 'Ljung-Box')

        Box-Ljung test

data:  fit$residuals
X-squared = 2.0107, df = 1, p-value = 0.1562
```

Fig. 21. Ljung ARIMA

```
> checkresiduals(fit)

        Ljung-Box test

data:  Residuals from ARIMA(0,0,4) with non-zero mean
Q* = 8.6883, df = 3, p-value = 0.03374

Model df: 5.   Total lags used: 8
```

Fig. 22. Residuals ARIMA

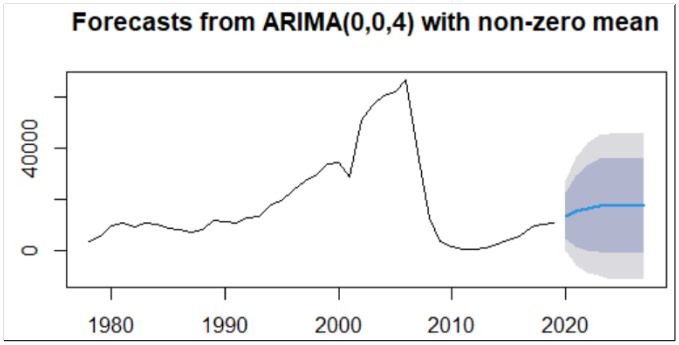

Fig. 23. Plot for ARIMA forecast

III. Logistic Regression

Logistic regression is a model that uses a statistical function to form a model for a dichotomous target attribute (dependent variable) (values can be either 1 or 0 , yes or no or etc). We cannot perform linear regression for a categorical variable as even if we plot a graph it will be a S shaped graph shifting from 0 to 1.

Child birth data set is used in this project to experiment on different test and logistic regression

A. Child Birth

Data Description:

This data set contains 16 attributes and 42 observations. We will pick our target variable to be "lowbwt" that is a binary target variable.

Exploratory Analysis:

```
Model Information:
Holt's method

Call:
 holt(y = dftimeseries_holt, h = 5)

  Smoothing parameters:
    alpha = 0.9999
    beta  = 0.4922

  Initial states:
    l = 6960.6255
    b = 967.5607

  sigma:  7894.782

     AIC     AICc      BIC
916.5910 918.2577 925.2794

Error measures:
                    ME     RMSE      MAE      MPE     MAPE     MASE      ACF1
Training set 5.943027 7509.436 4276.987 63.00985 87.12899 1.078075 0.1235899
```

Fig. 24. Summary for Holt's model

```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = dftimeseries, h = 2)

  Smoothing parameters:
    alpha = 0.9995

  Initial states:
    l = 3848.3497

  sigma:  7561.043

     AIC     AICc      BIC
911.1171 911.7487 916.3302

Error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE
Training set 165.2083 7378.822 3875.789 -7.771767 33.33161 0.9769476
                   ACF1
Training set 0.4218681

Forecasts:
     Point Forecast     Lo 80    Hi 80      Lo 95    Hi 95
2020       10783.75  1093.883 20473.62  -4035.623 25603.12
2021       10783.75 -2916.483 24483.98 -10168.948 31736.45
```

Fig. 25. Summary for SES

```
Series: dftimeseries
ARIMA(0,0,4) with non-zero mean

Coefficients:
         ma1     ma2     ma3     ma4       mean
      1.2055  1.0594  0.7716  0.5210  17464.631
s.e.  0.1816  0.2025  0.1621  0.1567   4373.654

sigma^2 estimated as 47419262:  log likelihood=-429.38
AIC=870.75   AICc=873.15   BIC=881.18

Training set error measures:
                    ME    RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set 196.2045 6463.29 4080.785 -55.55867 73.61107 1.02862 0.2112108
```

Fig. 26. Summary for ARIMA

We begin with the Chi - Square test that indicates the relationship with the categorical variable within the population.

Figure 27 presents the pearson chi-Square that is 0.007 for "mnocig" which means we reject the null hypothesis H0 and we accept the alternative. we say that "mnocig" variable will have impact on the target variable. So, we will use it for our regression model.

Figure 28 presents the pearson chi-Square that is 0.520 for "mage35" we say that "mage35" variable will not have impact on the target variable. So, we will not consider it for our regression model.

Similarly we verify the same with other variables too.

Figure 29 represents test for homogeneity which should pass. Based on mean its 0.93 which is greater than 0.05 . So, We can go with one way ANOVA test.

We consider null hypothesis as there is a difference in the variability. Here, we fail to reject the null hypothesis.

If we reject the null hypothesis then we go with robust test.

Figure 30 represents significance as .000 which is less than 0.05. So, we reject null hypothesis and accept the alternative hypothesis.

This means Length variable is helpful in predicting the target variable. There should be a variability in the mean.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	19.269[a]	7	.007
Likelihood Ratio	16.098	7	.024
Linear-by-Linear Association	.051	1	.821
N of Valid Cases	42		

a. 13 cells (81.3%) have expected count less than 5. The minimum expected count is .14.

Fig. 27. Chi Square for length variable

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.414[a]	1	.520		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.358	1	.549		
Fisher's Exact Test				.474	.474
Linear-by-Linear Association	.405	1	.525		
N of Valid Cases	42				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is .57.

b. Computed only for a 2x2 table

Fig. 28. Chi - Square for mage35 variable

If mean is same, then having two different groups in the consideration does not make any difference.

We performed logistic regression with the variable that had best relation like length, birthweight, Gestation, Headcirc, smoker, mnocig and mppwt. It can be observed that from fig 32 accuracy is 100 percent. This is not a good model. Accuracy cannot be 100 percent. This tells that the model is over fitting as we have less amount of data. So, to improve our model we create more attributes.

PCA:

This is a multivariate method that can be used to analyse on the data with values that are already inter related with each other. It extracts data as a set of components and identifies patterns with regards to the similarity. In generic terms, PCA

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Length	Based on Mean	.007	1	40	.933
	Based on Median	.006	1	40	.937
	Based on Median and with adjusted df	.006	1	39.996	.937
	Based on trimmed mean	.010	1	40	.920

Fig. 29. Test for homogeneity

ANOVA

Length

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	131.444	1	131.444	23.696	.000
Within Groups	221.889	40	5.547		
Total	353.333	41			

Fig. 30. ANOVA

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Length	-6.728	8076.720	.000	1	.999	.001
	Birthweight	-51.646	27927.570	.000	1	.999	.000
	Headcirc	6.675	1821.560	.000	1	.997	792.731
	Gestation	-5.768	6183.537	.000	1	.999	.003
	smoker	44.277	55292.365	.000	1	.999	1.695E+19
	mnocig	-.735	1009.570	.000	1	.999	.479
	mppwt	-.257	1836.613	.000	1	1.000	.773
	Constant	448.543	157331.777	.000	1	.998	6.304E+194

a. Variable(s) entered on step 1: Length, Birthweight, Headcirc, Gestation, smoker, mnocig, mppwt.

Fig. 31. Variables in equation

helps in summarizing the data values, which in turn will help in easy understanding of the visualization.

From figure fig 33 we can observe that 4th component can be considered as it has most of the relevance.

Visualizing a scree plot for the same will help in better understanding and picking the variables for the model. Fig **??** represents the component numbers that are to be considered above the elbow. And most importantly the difference must not be too high between the initial and final value.

After creating new factors with PCA, we will perform binary logistic regression.

Fig 35 indicates that we may consider the first four variables for our logistic model.

Fig 36 depicts the PCA accuracy as 95.2 % which is better than the previous 100% . It can be made more efficient by tuning the threshold that is the classification cutoff.

Fig 37 represents with the classification cutoff tuned to 0.4. It depends on what is the expected result from target variable.

Classification Table[a]

			Predicted		
			lowbwt		Percentage Correct
	Observed		0	1	
Step 1	lowbwt	0	36	0	100.0
		1	0	6	100.0
	Overall Percentage				100.0

a. The cut value is .500

Fig. 32. Initial accuracy as 100 percentage

Component Matrix[a,b]

	Component			
	1	2	3	4
Length	.840	-.140	.099	-.501
Birthweight	.008	-.282	.689	.612
Headcirc	.183	.214	.929	.043
Gestation	.556	-.077	.600	-.455
mage	-.418	.838	.019	-.326
mnocig	.674	.312	.016	.669
mheight	.754	.573	-.260	.097
mppwt	.483	.849	.093	.093
fage	-.461	.855	.141	-.161
fedyrs	-.657	.240	.700	.109
fnocig	.907	.194	-.327	.178
fheight	.799	-.255	.472	-.260

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

b. Only cases for which lowbwt = 1 are used in the analysis phase.

Fig. 33. Principle component analysis

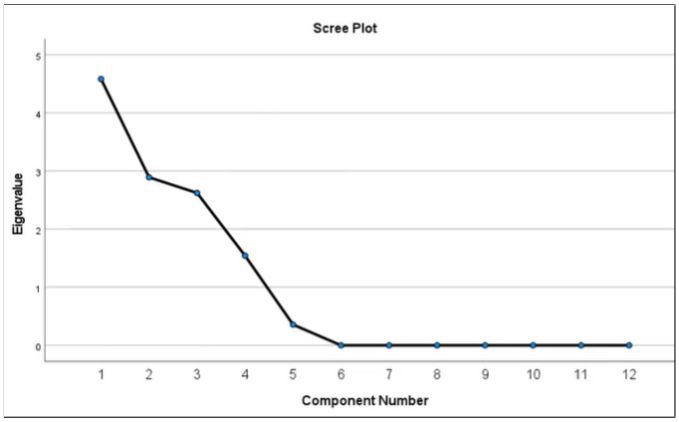

Fig. 34. Scree plot

Like if the child birth weight prediction is important then we need to set threshold one by one.

Still the model can be made better by increasing the threshold. So, setting the classification cutoff to 0.7 figure 38 represents the the maximum of true positive and less of false negative.

The same model was executed without PCA and the accuracy can be viewed in figure 39 which gives 100 % accuracy and this is a over fitted model with out PCA.

Total Variance Explained[a]

Component	Initial Eigenvalues Total	% of Variance	Cumulative %	Extraction Sums of Squared Loadings Total	% of Variance	Cumulative %
1	4.585	38.204	38.204	4.585	38.204	38.204
2	2.892	24.103	62.307	2.892	24.103	62.307
3	2.623	21.859	84.166	2.623	21.859	84.166
4	1.543	12.859	97.025	1.543	12.859	97.025
5	.357	2.975	100.000			
6	1.382E-15	1.152E-14	100.000			
7	5.502E-16	4.585E-15	100.000			
8	2.221E-16	1.851E-15	100.000			
9	8.815E-17	7.346E-16	100.000			
10	-1.305E-16	-1.088E-15	100.000			
11	-2.530E-16	-2.108E-15	100.000			
12	-7.205E-16	-6.004E-15	100.000			

Extraction Method: Principal Component Analysis.

a. Only cases for which lowbwt = 1 are used in the analysis phase.

Fig. 35. Test of variance

Classification Table[a]

			Predicted lowbwt 0	1	Percentage Correct
Step 1	Observed lowbwt	0	35	1	97.2
		1	1	5	83.3
	Overall Percentage				95.2

a. The cut value is .500

Fig. 36. PCA Accuracy

Classification Table[a]

			Predicted lowbwt 0	1	Percentage Correct
Step 1	Observed lowbwt	0	34	2	94.4
		1	1	5	83.3
	Overall Percentage				92.9

a. The cut value is .400

Fig. 37. with threshold of 0.4

Classification Table[a]

			Predicted lowbwt 0	1	Percentage Correct
Step 1	Observed lowbwt	0	36	0	100.0
		1	3	3	50.0
	Overall Percentage				92.9

a. The cut value is .700

Fig. 38. with threshold of 0.7

Classification Table[a]

			Predicted lowbwt 0	1	Percentage Correct
Step 1	Observed lowbwt	0	36	0	100.0
		1	0	6	100.0
	Overall Percentage				100.0

a. The cut value is .500

Fig. 39. Accuracy without PCA

IV. Conclusions and Future Work

With the above experiments made on the time series and Logistic regression, we can conclude that for both the time series data sets ARIMA was the model with best accuracy and RMSE. In fact, time series helps to forecast with a good accuracy by analysing the historic data.

In logistic regression, use of PCA helped in extracting the best model that could be brought with less number of attributes in the given.

References

[1] *Logistic Regression https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis*

[2] *ARIMA Documentation*, *https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/*,

[3] *Time Series , Pattern Extraction for Time Series Classification*, *https://link.springer.com/chapter/10.1007/3540447946*$_1$0