LOGISTIC REGRESSION ANDTIME SERIES ANALYSIS STATISTICS-CA-2

Iswarya Yogeashwaran
M.Sc in Data Analytics
National college of Ireland
X20155034

Abstract— **This project is to evaluate and analyze the time series and logistic regression. The time series analysis has estimatedby using two datasets for getting best fit model among the three different appropriately used models. The logistic regression is performed in the dataset given to estimate the most significant predictor variable and independent variables by using relevant dimensional reduction methodology. The reasons for selectingbest fit models are explained clearly.**

Keywords- timeseries, model, RMSE, logistic regression.

I. INTRODUCTION

A. TIME SERIES ANALYSIS

A time series data set is a range of multiple values thatan attribute fluctuates over time. Time series analysis is the method for dealing with time series data and evaluating the insights. This method utilizes seasonality, trend, and patterns of time series data, which aids in forecasting future events.

B. LOGISTIC REGRESSION

The statistical methodology of logistic regression is used to predict the result of the response variable (binary variable). For logistic regression, the independent variables can be continuous or categorical, but the response variable must be dichotomous or nominal or ordinal. The equation for logistic regression is,

$$\hat{Y} = \frac{e^{b_0+b_1X_1+\cdots+b_kX_k}}{e^{b_0+b_1X_1+\cdots+b_kX_k}+1}$$

Fig. 1. Equation of Logistic Regression

II. OBJECTIVE

A. TIME SERIES ANALYSIS

This analysis is used to find the optimum model by checkingappropriate test from the three model fits for the given two datasets such as overseas trips and new home registration. Forecasting the next three periods of the optimum model.

B. LOGISTIC REGRESSION

In this analysis, the appropriate dimensional reduction method is used to find the significant dichotomous variable among the 17 variables and the reason for the selection of the significant model.

III. METHODOLOGY

A. TIME SERIES ANALYSIS

1) DATA DICTIONARY*:* The dataset used in this time series analysis are "The overseas trips" and "The new home registration". "The overseas trips" prediction will be forecasted by the overseas trips by the non-residents to Ireland from the period of 2012 of quarter 1 to 2019 of quarter 4. "The overseas trips" is a quarterly time-period dataset. This dataset consists of two columns such as "year" and "Trips.Thousands", which has 32 observations. "The new home registration" prediction will be forecasted by the registration from the year 1978 tothe year 2019. "The new home registration" is an annual time-period dataset. This dataset contains only two columns namely"i.year" and "new home registration", that has 42 rows.

1. DATA CLEANING AND PRE-PROCESSING*:* The two datasets are already cleaned and pre-processed. These

datasetsare used as .csv file format before loading into R studio. TheR programming is used in these two datasets to estimate the models of the time series.

2) ASSUMPTIONS VERIFICATION*:* The required librariesare loaded into R studio. The loaded libraries are fpp2, ggplot2and tseries for both the datasets. The fpp2 package is used for forecasting of time series, ggplot2 is used for data visualizationand tseries is utilized for the time series analysis. The datasets in .csv file format are loaded as dataframe.

3) DATASET-1*: Overseas Trips:* The "Trips.Thousands" column from the dataframe is converted into object using ts() function. The time series data is visualized to see the insights using plot function (Figure-2).

- COMPONENTS OF RAW TIME SERIES

The components of a raw time series are the numerous rationale that affect the values of an assessment in a time series. The components are of four types: 1. Trend 2. Seasonal Changes 3. Cyclic Changes 4. Unpredictable or irregular movement Seasonal and cyclical variations are changes that occur on a regular basis or are short-term changes. This is observed that this time series values have trend and seasonality. The Linear trend is present in this time series, which is justified by plotting abline function, which is seen in the figure-3. The abline function is plotting mean to the data. The seasonality is checkedby using seasonal plot which is visualized by the helpof ggplot package in the figure-4. The level of the time series is given in the figure-5. The lines represent the mean of each quarter period. In the figure-6,the autoplot isa function that provides better default graphics for various data objects.

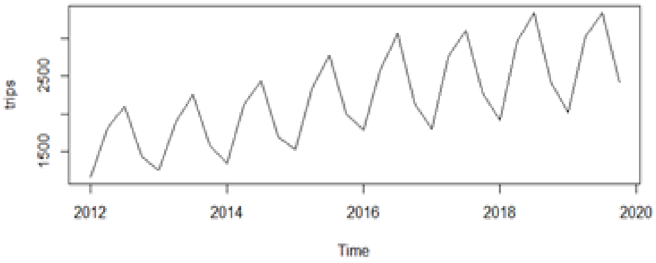


Fig. 2. Plot of Raw data

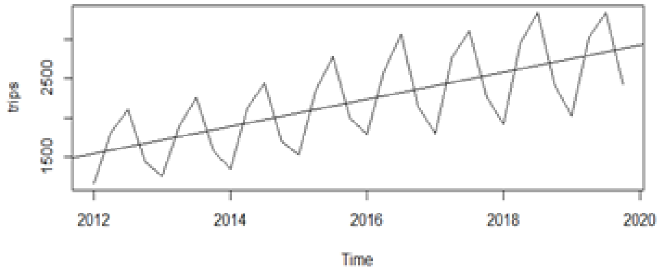


Fig. 3. Trend of Raw data

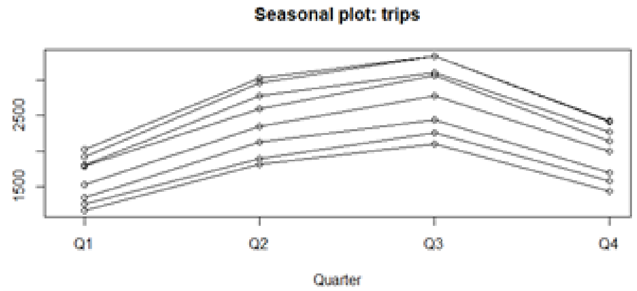


Fig. 4. Seasonal Plot

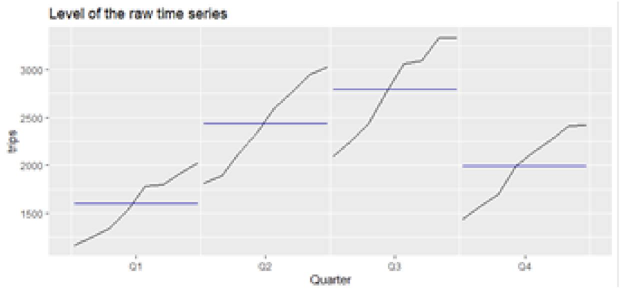


Fig-5 Level of time series

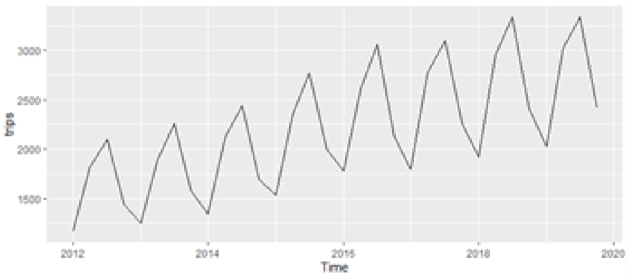


Fig-6 Auto-plot of time series

- SEASONAL DECOMPOSITION:

The time series data has seasonal effect, which decomposed into seasonality, trend, and random error. The decomposition of seasonality is of two types: additive and multiplicative. The additive means seasonality will be constant, it is seen in the figure-7. The multiplicative means there will be growth in the time series when seasonality changes, it is seen in the figure-8.

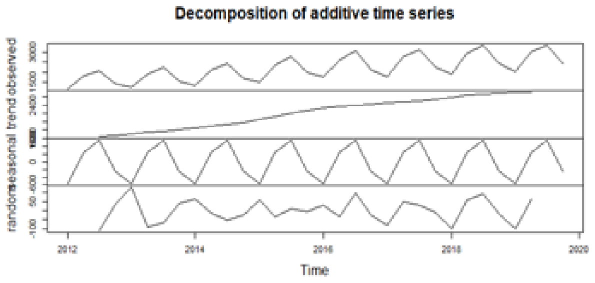


Figure-7 Additive decomposition

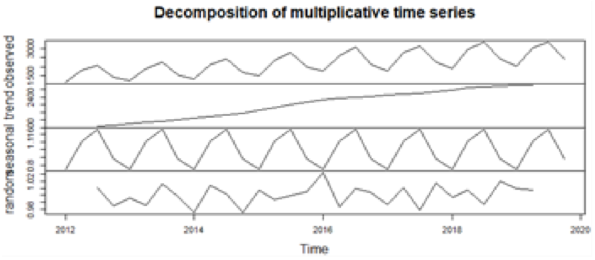


Figure-8 Multiplicative decomposition

- MODEL FITTING:

In this time series, the four models used to evaluate are **seasonal naive, Holt winters seasonal method, ETS function model and Arima model**. The plot for seasonal naive is shown in the Figure-9. The RMSE value for seasonal naive is 176.6505 in shown in the figure-10.

This time series has both seasonality and trend. So, Holt winters seasonal model is used for forecasting. Holt winters has both additive and multiplicative model. The additive model fits well in the plot, it can be seen in the figure-11. The Holt-Winters' additive model is utilized to forecast, it is shown in figure-12. The RMSE valuefor Holt-Winter's model is 76.60357. The Holt-Winter's model has AIC value of 406.5767(figure-13 shows the accuracy and AIC Value). The ETS function is a functionused to automatically get the model. Therefore, ETS gives (M,A,M) model. This ETS function has RMSE value of 54.698 and AIC value of 378.8267(Figure-14).

In Arima model, the data should be stationary. This means the data should have constant mean, variance,and covariance. The stationary data should never have trend and seasonality. Therefore, this time series is not stationary as it has trend and seasonality. So, diff functionis used to change non-stationary into stationary. **Augment Dickey Fuller's test** is used to check the stationarity. Afterdoing difference functionality, p value has 0.01 value, which is significant, it is shown in the figure-15. So, it rejects the null hypothesis and accepts the alternative hypothesis. hypothesis, which means it is stationary.

Box-Luang test is used to check the auto correlation of residuals or white noise present in the data. It gives significantly the model does not have lack of fit. Becausep-value is not significant, that accepts the null hypothesis. Arima model (0,1,1) is fitted. The ACF (Auto-correlation Function) has no lags in the model. The data is normally distributed. These can be seen in the figure-16. The RMSE value for Arima model is 67.54754, it is shown in the figure-17. The AIC value for exponential smoothing models is better for ETS function which gives M,A,M model(378.8267) than Holt winter's smoothing model. The RMSE value is lowest for ETS function of M,A,M model, which is 54.698 comparing to other three models. The forecast for ETS function(M,A,M) model is given in the figure-19.

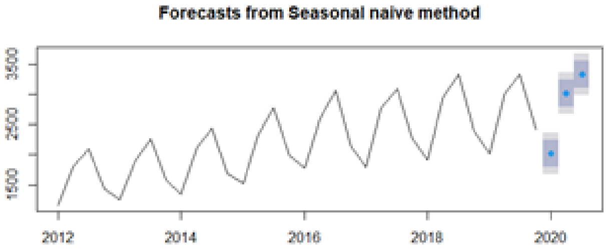
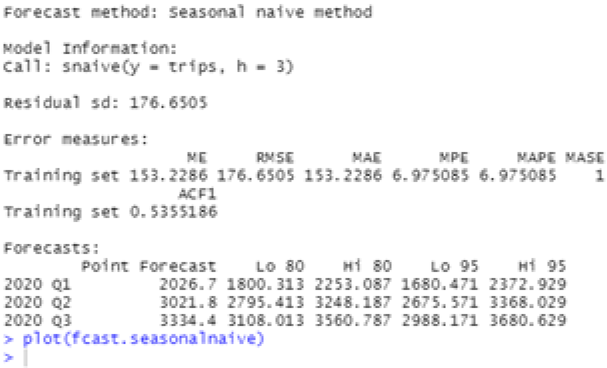


Figure-9 Seasonal Naïve forecast plot



```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = trips, h = 3)

Residual sd: 176.6505

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

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1         2026.7 1800.313 2253.087 1680.471 2372.929
2020 Q2         3021.8 2795.413 3248.187 2675.571 3368.029
2020 Q3         3334.4 3108.013 3560.787 2988.171 3680.629
> plot(fcast.seasonalnaive)
> 
```

Figure-10 Accuracy of seasonal naïve model

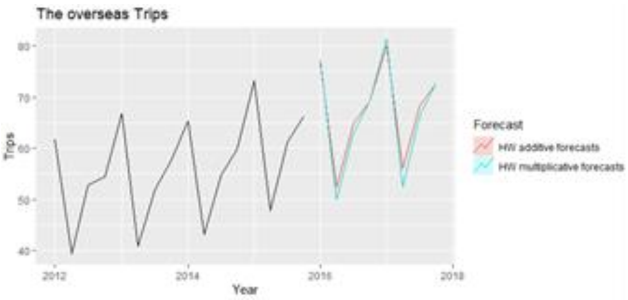


Figure-11 Smoothing model- moving average.

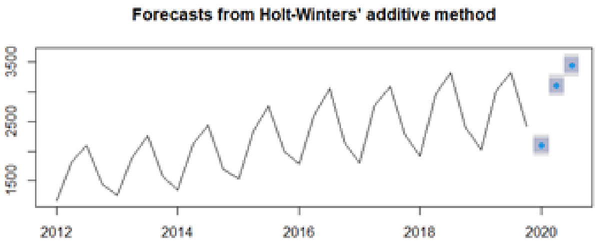
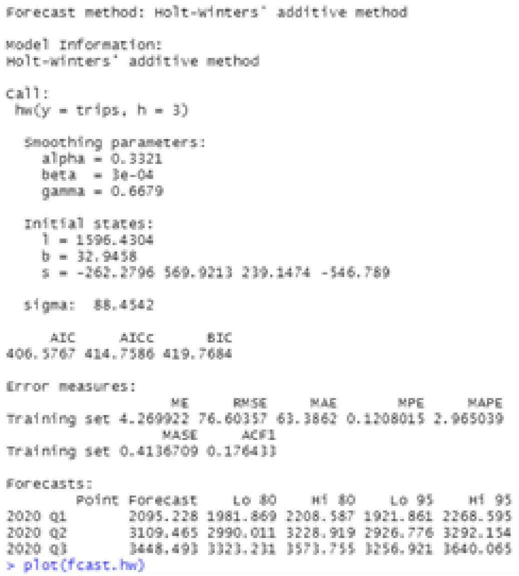
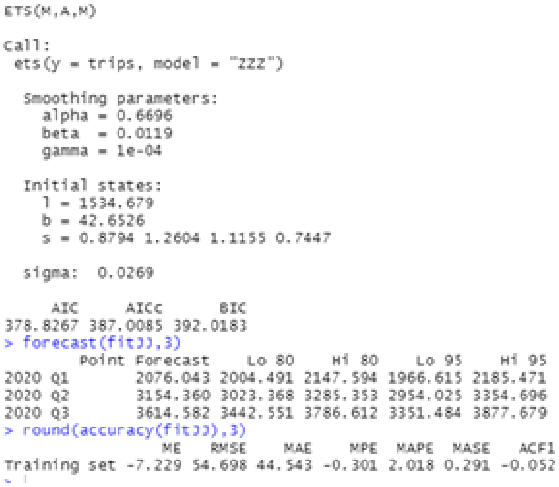


Figure-12 Holt-Winter's additive method



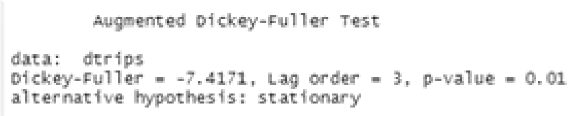
```
Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

Call:
 hw(y = trips, h = 3)

  Smoothing parameters:
    alpha = 0.3321
    beta  = 3e-04
    gamma = 0.6679

  Initial states:
    l = 1596.4304
    b = 32.9458
    s = -262.2796 569.9213 239.1474 -546.789

  sigma:  88.4542

     AIC     AICc      BIC
406.5767 414.7586 419.7684

Error measures:
                   ME     RMSE     MAE       MPE     MAPE
Training set 4.269922 76.60357 63.3862 0.1208015 2.965039
                  MASE     ACF1
Training set 0.4136709 0.176433

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2095.228 1981.869 2208.587 1921.861 2268.595
2020 Q2       3109.465 2990.011 3228.919 2926.776 3292.154
2020 Q3       3448.493 3323.231 3573.755 3256.921 3640.065
> plot(fcast.hw)
```

Figure-13 Accuracy of Holt-winter's additive



```
ETS(M,A,M)

Call:
 ets(y = trips, 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
> forecast(fitJJ,3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2076.043 2004.491 2147.594 1966.615 2185.471
2020 Q2       3154.360 3023.368 3285.353 2954.025 3354.696
2020 Q3       3614.582 3442.551 3786.612 3351.484 3877.679
> round(accuracy(fitJJ),3)
                 ME   RMSE    MAE    MPE  MAPE  MASE   ACF1
Training set -7.229 54.698 44.543 -0.301 2.018 0.291 -0.052
```

Figure-14 Accuracy of ETS function



```
        Augmented Dickey-Fuller Test

data:  dtrips
Dickey-Fuller = -7.4171, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

Figure-15 Augmented Dickey-Fuller Test

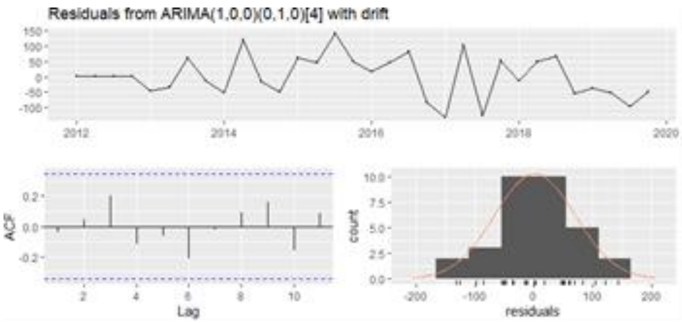
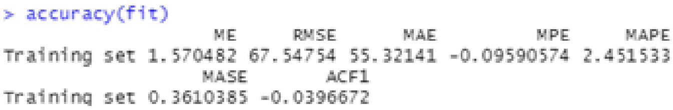


Figure-16 Residuals of Arima model



```
> accuracy(fit)
                   ME     RMSE      MAE         MPE     MAPE
Training set 1.570482 67.54754 55.32141 -0.09590574 2.451533
                  MASE       ACF1
Training set 0.3610385 -0.0396672
```

Figure-17 Accuracy of Arima model

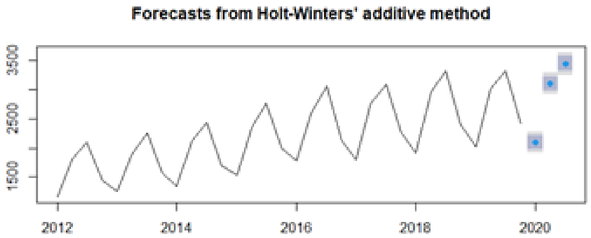
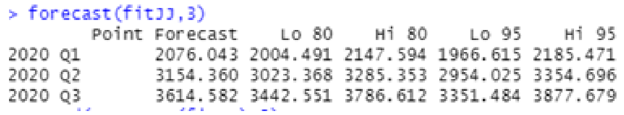


Figure-18 Forecast Plot from ETS Function



```
> forecast(fitJJ,3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2076.043 2004.491 2147.594 1966.615 2185.471
2020 Q2       3154.360 3023.368 3285.353 2954.025 3354.696
2020 Q3       3614.582 3442.551 3786.612 3351.484 3877.679
```

Figure-19 Forecast from Arima model

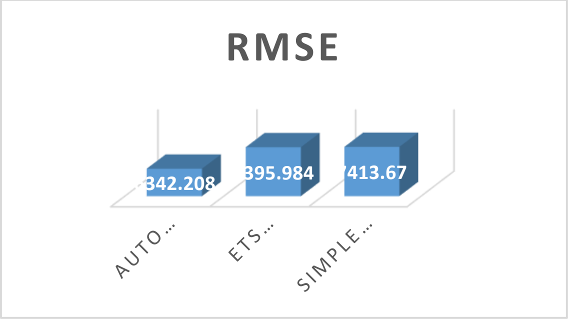


Figure-20 RMSE values

The ETS function is the efficient model in this dataset.

5.DATASET-2: THE NEW HOME REGISTRATION:

The “newhome registration” column is converted into ts object. This time series is plotted, which can be seen in the figure.

- **Components of Raw series**:

The abline function is used to check the trend and seasonality plot is plotted to checkthe seasonality. Th figure represents that there is no trendin the data and seasonality plot function provides the datahas no seasonality.(Figure-22).

The Smoothing time series is a moving average, which is smoothed by using moving average value 3 and 5, where 5 is better and not over or under smoothed. The moving average is used to smooth the raw time series.(Figure-23)

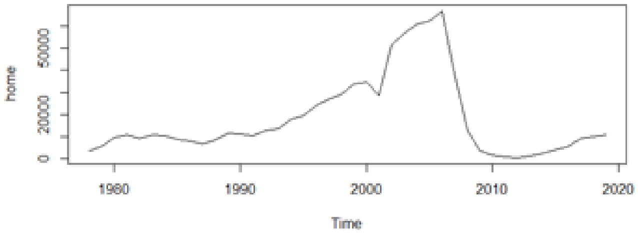


Figure-21 Plot of raw data

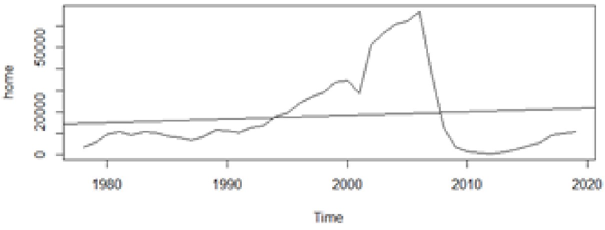


Figure-22 Trend of the time series

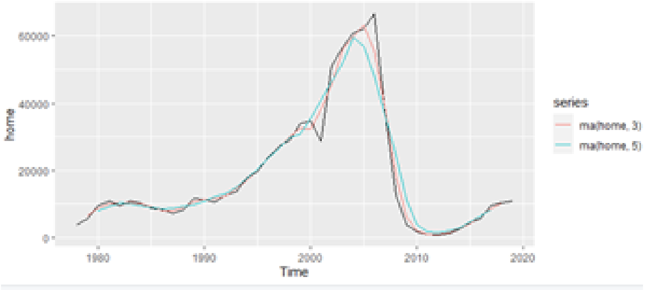
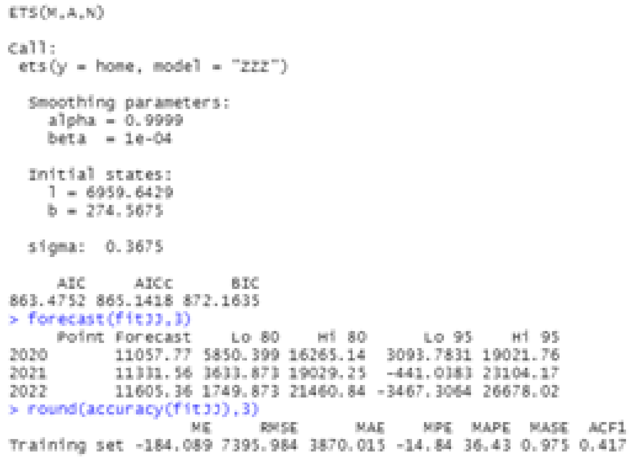


Figure-23 Smoothing time series.

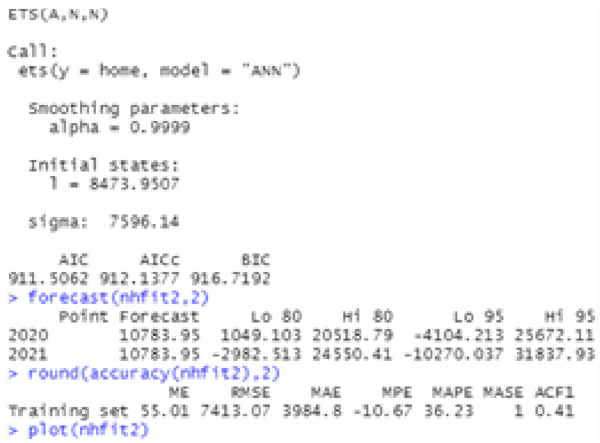
- **Model fitting**:

In this time series, Arima model, ETS function of exponential smoothing model and Simple exponential smoothing model are used. The ETS function is used to automatically get model from the software, which has RMSE value of 7395.984 and AIC value of 863.4752. The simple exponential smoothing model is fitted and has RMSE value of 7413.67 and AIC value of 864.86.



```
ETS(M,A,N)

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

  Smoothing parameters:
    alpha = 0.9999
    beta  = 1e-04

  Initial states:
    l = 6959.6429
    b = 274.5675

  sigma:  0.3675

     AIC     AICc      BIC
863.4752 865.1418 872.1635
> forecast(fitJJ,3)
     Point Forecast    Lo 80    Hi 80      Lo 95    Hi 95
2020       11057.77 5850.399 16265.14  3093.7831 19021.76
2021       11331.56 3633.873 19029.25  -441.0383 23104.17
2022       11605.36 1749.873 21460.84 -3467.3064 26678.02
> round(accuracy(fitJJ),3)
                   ME     RMSE      MAE    MPE  MAPE  MASE  ACF1
Training set -184.089 7395.984 3870.015 -14.84 36.43 0.975 0.417
```

Figure-24 ETS function accuracy and forecast

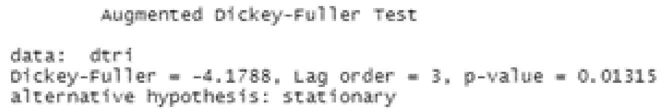


```
ETS(A,N,N)

Call:
 ets(y = home, model = "ANN")

  Smoothing parameters:
    alpha = 0.9999

  Initial states:
    l = 8473.9507

  sigma:  7596.14

     AIC     AICc      BIC
911.5062 912.1377 916.7192
> forecast(nhfit2,2)
     Point Forecast     Lo 80    Hi 80      Lo 95    Hi 95
2020       10783.95  1049.103 20518.79  -4104.213 25672.11
2021       10783.95 -2982.513 24550.41 -10270.037 31837.93
> round(accuracy(nhfit2),2)
                ME    RMSE    MAE    MPE  MAPE MASE ACF1
Training set 55.01 7413.07 3984.8 -10.67 36.23    1 0.41
> plot(nhfit2)
```

Figure-25 ETS function accuracy and forecast

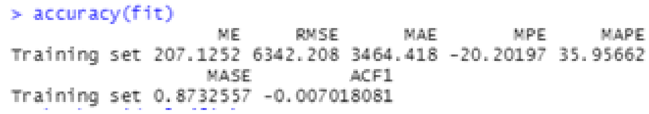
The **ADF (Augment Dickey Fullers test)** is used to check stationary(Figure-26). It shows that it is not stationary. So, differenceis taken for the data by using diff function. Again, ADF test is to be checked and it provides p value of 0.01, which is significant and rejects null hypothesis and accepts alternative hypothesis. It is now stationary. The Arima model (0,1,2) is to be fitted and it gives no lags in ACF. The residuals are normally distributed. The **Luang Box test** is used to check autocorrelation. The RMSE value for Arima model is 6342.208.(figure-27).

The AIC Value of ETS function is 863.4752, which is lowest compared to simple exponential smoothing, which has 911.5062. The RMSE value of Arima modelis 6342.208, which is lowest value compared to other three models. Therefore, the forecasting of three period are provided in the figure-28.



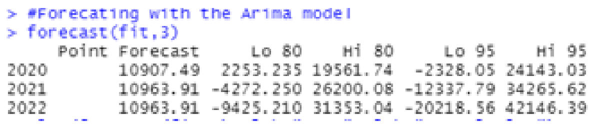
```
        Augmented Dickey-Fuller Test

data:  dtri
Dickey-Fuller = -4.1788, Lag order = 3, p-value = 0.01315
alternative hypothesis: stationary
```

Figure-26 Augmented Dickey-Fuller Test



```
> accuracy(fit)
                   ME     RMSE      MAE       MPE     MAPE
Training set 207.1252 6342.208 3464.418 -20.20197 35.95662
                  MASE         ACF1
Training set 0.8732557 -0.007018081
```

Figure-27 Accuracy of Arima model



```
> #Forecating with the Arima model
> forecast(fit,3)
     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2020       10907.49  2253.235 19561.74  -2328.05 24143.03
2021       10963.91 -4272.250 26200.08 -12337.79 34265.62
2022       10963.91 -9425.210 31353.04 -20218.56 42146.39
```

Figure-28 Forecast of Arima model

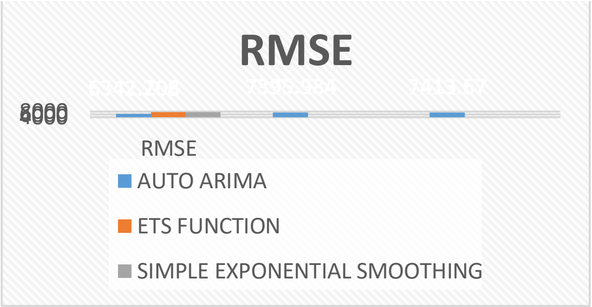


Figure- 29 RMSE value

The RMSE value is efficient for Auto Arima model.

B. LOGISTIC REGRESSION

1) DATA DICTIONARY: The dataset provided has contains17 columns and rows.

2) DATA CLEANING AND TRANSFORMATION: Data has no null values. The Multicollinearity is to be checked and no variables are highly correlated with one another. This proves by the figure-29.

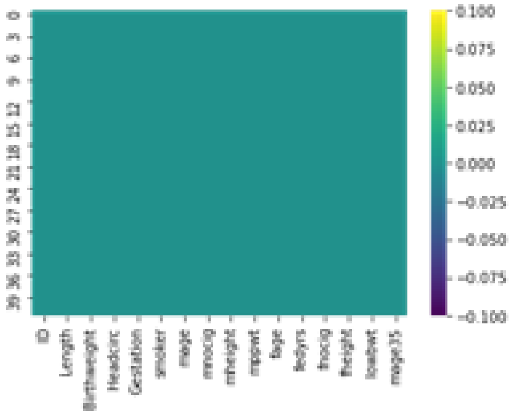
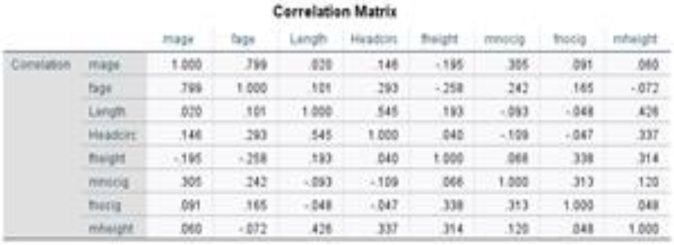


Figure-30 Missing values



Correlation Matrix

		mage	fage	Length	Headcirc	fheight	mnocig	fnocig	mheight
Correlation	mage	1.000	.799	.020	.146	-.195	.305	.091	.060
	fage	.799	1.000	.101	.293	-.258	.242	.165	-.072
	Length	.020	.101	1.000	.545	.193	-.093	-.048	.426
	Headcirc	.146	.293	.545	1.000	.040	-.109	-.047	.337
	fheight	-.195	-.258	.193	.040	1.000	.066	.338	.314
	mnocig	.305	.242	-.093	-.109	.066	1.000	.313	.120
	fnocig	.091	.165	-.048	-.047	.338	.313	1.000	.048
	mheight	.060	-.072	.426	.337	.314	.120	.048	1.000

Figure-31 Correlation Matrix

3). METHODOLOGY:

- ASSUMPTIONS:

1. SAMPLE SIZE:
   The sample size should be taken according to the samples in each section. Two variables are fitted in this logistic regression problem. The assumption is proved.
2. OUTLIERS:
   To remove all outliers, the outlier capping process is

being used to impute all outliers with quantiles. In this dataset, the value above the quartile is limited. After all the outliers have been imputed using the outlier capping methods, the tiny proportion of remaining outliers are removed, and one column is dropped to reduce data loss by dropping all the outliers in all columns.

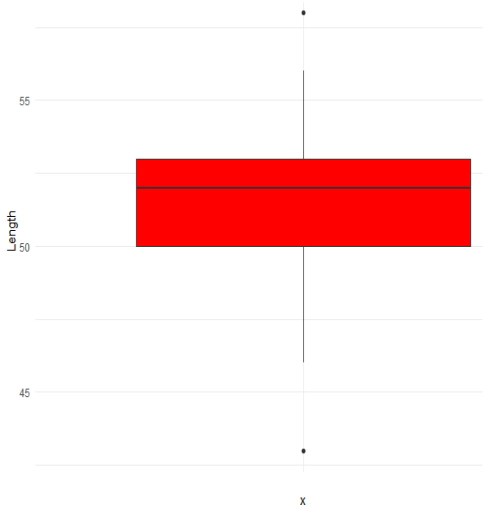


Figure-32 Checking outliers before imputation.

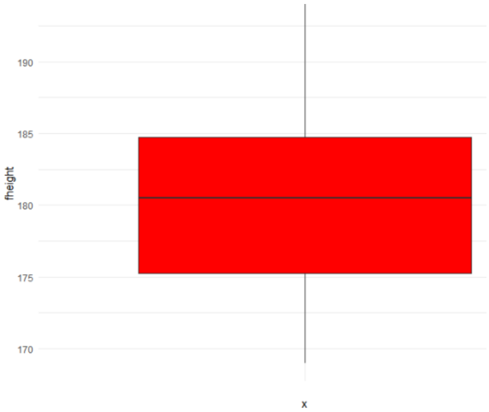


Figure-33 Checking Outliers after imputation.

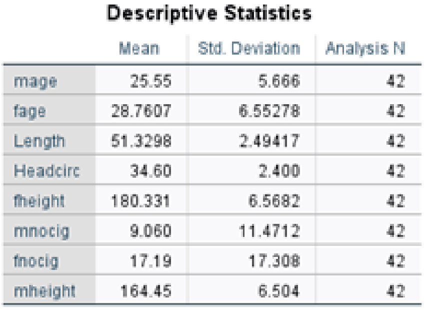
3. MULTICOLLINEARITY:
There should no high relationship between the variables. The correlation is checked, in the figure-29, that there is no relationship between the variables. Therefore, the assumption is proved.

4. MUTUALLY EXCLUSIVE:
The target variable – low birth weight is a binary variable, which is mutually exclusive in this childbirth dataset. It is 0 and 1. The 0 value means low birth weight and 1 value means no low birth weight. Therefore, this assumption is proved.

4) ANALYSIS PROCEDURE:

- Import the Data in SPSS.
- Navigate to Analyze¡- Regression ¡- Binary LogisticRegression.
- Drag and drop the variable “low birth weight” in targetvariable box.
- Drag and drop the other relevant variables in the independent variable box.
- Click KMO and Bartlett’s test in Descriptive statisticsoption.
- Click Fixed 2 in the extraction option.

The independent variables used in the analysis provided in Descriptive statistics(figure-34).

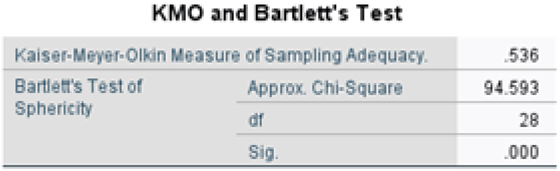


Descriptive Statistics

	Mean	Std. Deviation	Analysis N
mage	25.55	5.666	42
fage	28.7607	6.55278	42
Length	51.3298	2.49417	42
Headcirc	34.60	2.400	42
fheight	180.331	6.5682	42
mnocig	9.060	11.4712	42
fnocig	17.19	17.308	42
mheight	164.45	6.504	42

Figure-34 Descriptive statistics

The KMO and Bartlett’s test is checked for PCA Dimensional reduction method (factor analysis) to select two significant independent variables to fit into the model. The Accuracy for KMO and Bartlett’s test provide 0.536,which should be greater than 0.5 and also it is significantby 0.0 p- value(Figure-35).



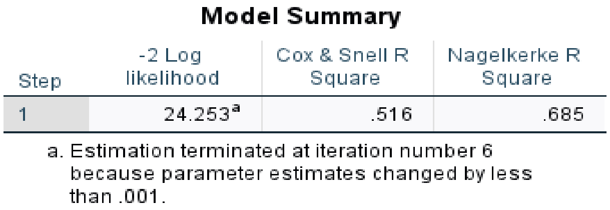
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.536
Bartlett's Test of Sphericity	Approx. Chi-Square	94.593
	df	28
	Sig.	.000

Figure-35 KMO and Bartlett’s Test

5)RESULTS INTERPRETATION:

The forward method is used to fit models. In the figure-36, Cox Snell R Square and Nagelkerke R square test value ranges 0.516 and 0.685. This value should be greater than 0.5.



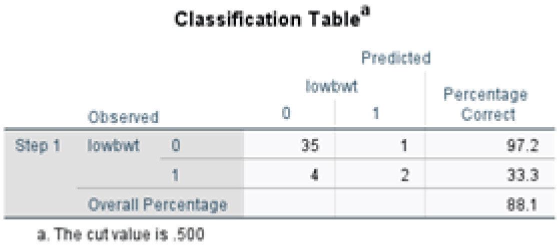
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24.253[a]	.516	.685

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Figure-36 Model Summary

In the figure-37, the overall prediction table obtained for the classification table is 88.1 and cut off value is 0.500 and itfalls in the group 1.



Classification Table[a]

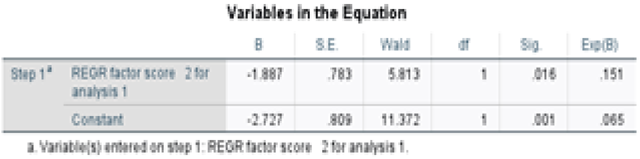
			Predicted		
			lowbwt		Percentage Correct
	Observed		0	1	
Step 1	lowbwt	0	35	1	97.2
		1	4	2	33.3
	Overall Percentage				88.1

a. The cut value is .500

Figure-37 Classification Table

In the figure-38, it shows that the variables are to be entered into the model are significant.

In the figure-39, Hosmer and Lemeshow Test showed chi-square value of 0.5446, which means the predictor variable is significant. From the above results, the model for predicting the lower birth weight of the child has proved to be classified as 88.1%.

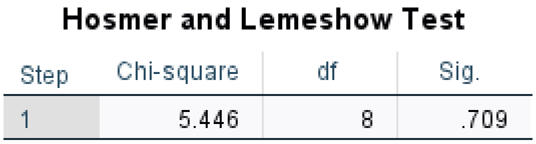


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 2 for analysis 1	-1.887	.783	5.813	1	.016	.151
	Constant	-2.727	.809	11.372	1	.001	.065

a. Variable(s) entered on step 1: REGR factor score 2 for analysis 1.

Figure-38 Variables in the Equation



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.446	8	.709

Figure-39 Hosmer and Leme show Test.

IV. CONCLUSION

The logistic regression for predicting the low birth weight of the child has proved that the model is fitted with accuracy 88.1%. The predictor variable is significant variable to predict. In the time series, “Overseas Trips” dataset has the best fit model as ETS function of Simple Exponential Smoothing with RMSE value of 54. In “new home registration” dataset, Arima model is the best fitted model with high accuracy (RMSE value of 6342.208). This concludes that these two models forecasted the next three periods efficiently.