TABA Statistics – Time Series & Logistic Regression

MSc. Data Analytics - Jan 2021

Sanket Sonu

x19206071@student.ncirl.ie

Prof: Tony Delaney

Abstract: **This project paper aims to demonstrate 2 important Statistic methods, Time Series and Logistic Regression. This project is executed on both R and IBM SPSS. Time Series is a technique to find the trend, patterns etc in a series of data based on time. Time Series has been implemented in the R programming language. Logistic Regression is a classification model, Binary logistic regression has been used here, this means the dependent variable has 2 values: True/False, Yes/No, etc. Logistic Regression is implemented on IBM SPSS.**

I. TIME SERIES

Time Series is a statistical technique which is used to find trends in a series of data based on time interval. Time Series data simply means that data is in series with respect to time.

There are three types of time series data:

1. Time Series data: It is a set of observations on the values that a variable takes at different time.

2. Cross-Sectional data: It is the data of the same point in time when data of one or more variable are collected.

3. Pooled data: Pooled data is a combination of both Time Series and Cross-Sectional data.

For this project, we are using two time-series data to forecast. R programming language has been used for both Time Series datasets.

A. "OverseasTrips":

This is seasonal time series data. The data contains trips to Ireland by non-residents. The time is from Quarter 1 of 2012 to Quarter 4 of 2019. Each year is divided into 4 quarters, i.e., Q1, Q2, Q3, and Q4.

First, a time-series object has been created using the ts() function and below (Fig. 1) shows the time series plot used for analysing patterns. This pattern shows the behaviour of time series over a period. The past pattern can be used for selecting the best forecasting method.

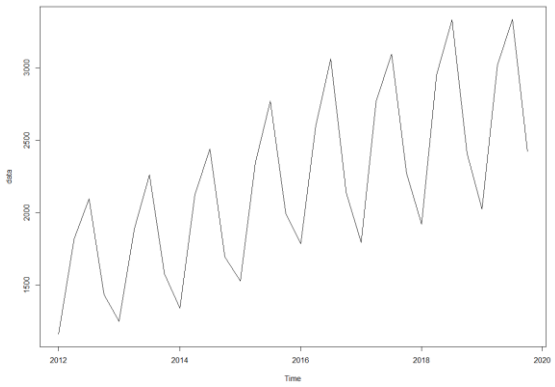

Fig. 1. Time Series Plot

Above (Fig. 1) shows a constant pattern in increasing order. This pattern can be termed variability, and it is constant over time. This time-series data has patterns like trend and seasonality, an increasing trend can be observed from this plot.

(Fig. 2) Shows the seasonal plot. This is a Seasonal Time Series data, which is based in term of 4 Quarters of years.

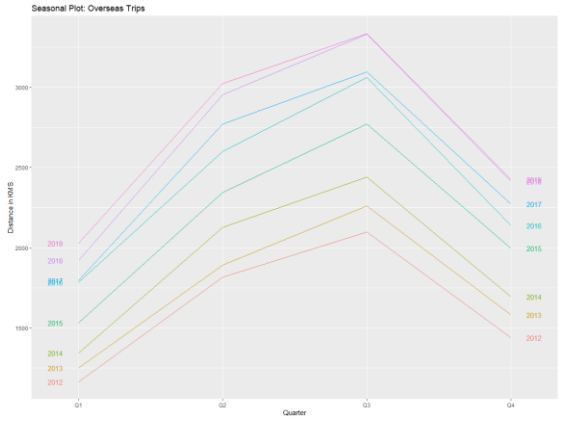

Fig. 2. Seasonal Plot

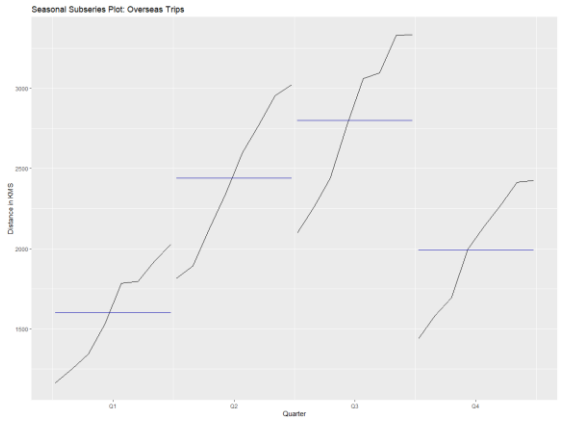

Fig. 3. Seasonal Subseries Plot

(Fig. 3) Shows the Seasonal Subseries Plot. This shows data for each season that are collected. The horizontal line shows the mean of each quarter. From the above Season Subseries Plot, it can be observed that mean of Quarter 3 is the highest among others and the mean of Quarter 1 is the lowest.

This time-series data is Seasonal can be decomposed into components. We are using 2 decomposition methods here: 1st is normal decompose () and 2nd is STL Decomposition.

(Fig. 4) Shows the decomposition of this seasonal time series data. From the below fig, it clearly observed that it is divided into 4 components, such as random, seasonal, trend, and observed.

The random component demonstrates the random errors obtained. Random errors are more at the start and end of the dataset. The seasonal component demonstrates the constant seasonality, it shows trips are less at starting and end of the year, i.e., Q1 and Q4. The trend component demonstrates the increasing trend in our dataset concerning time. The observed component demonstrates the time series plot, which shows increasing in pattern.

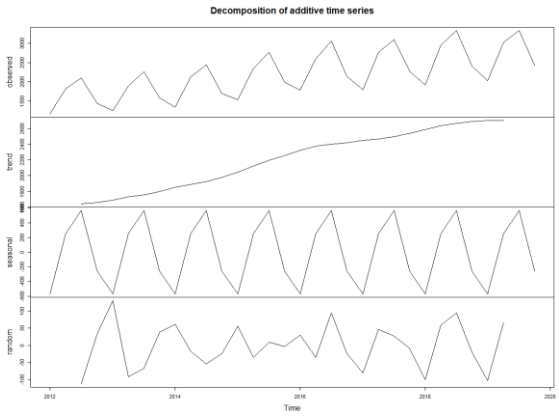

Fig. 4. Decompose of additive Time Series.

(Fig. 5) Shows the decomposition of this seasonal time series data. This is a robust method. Previous Classic decomposition had some drawbacks and to overcome those drawbacks, STL has been implemented. STL can handle every type of seasonality, not only days and months data. User can control the rate of change of Seasonal component. User can also control trend's smoothness. Because of its robustness to outliers, occasional random observations will not have an impact on the trend cycle and seasonality.

STL decomposition produces 4 components, such as remainder, trend, seasonal, and data.

The vertical bar on the right side of each component demonstrates a comparison of the magnitudes of each component.

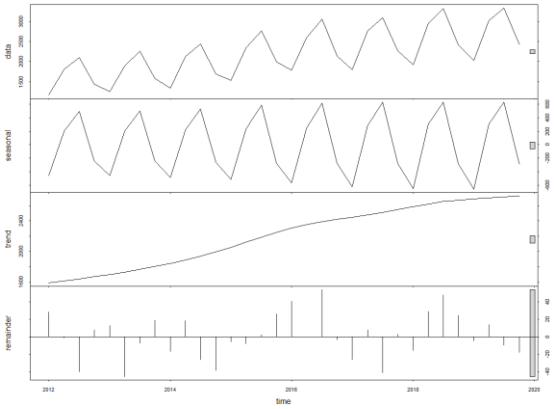

Fig 5. STL Decomposition

The trend component demonstrates the increasing trend in our dataset with respect to time. The seasonal component demonstrates the constant seasonality, it shows trips are less at starting and end of the year, i.e., Q1 and Q4. The data component is our actual data time series plot, which is increasing in pattern. The remainder plot shows the residuals that are generated from trend and seasonal fit. STL is robust to outliers, occasional random observations will not show impact on trend and seasonality, however, those will affect this Remainder component.

Next, data is spitting into train and test. The first six years, i.e., Q1 2012 to Q4 2017 will be used as train and the rest 2 years, i.e., Q1 2018 to Q4 2019 will be used as test data.

A. Forecast:

Next, using Forecast Library, time-series data can be predicted. For this model, decomposed data is used and the method as 'rwdrift'. The STL decomposed data used here is log10(train_data).

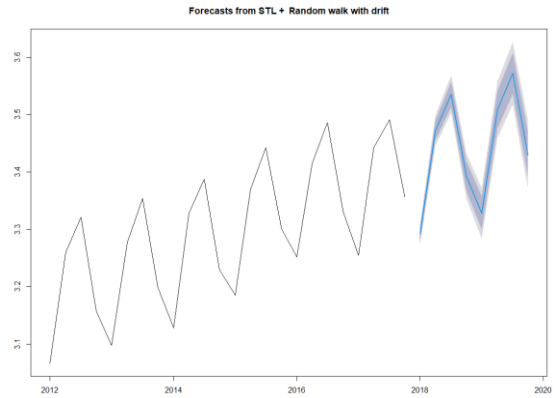

Fig. 6. Forecast from STL + Random walk with drift.

The grey part (Fig. 6) shows the confidence interval, and the blue line is the average of the forecast. It shows, a forecast can go into grey area as well.

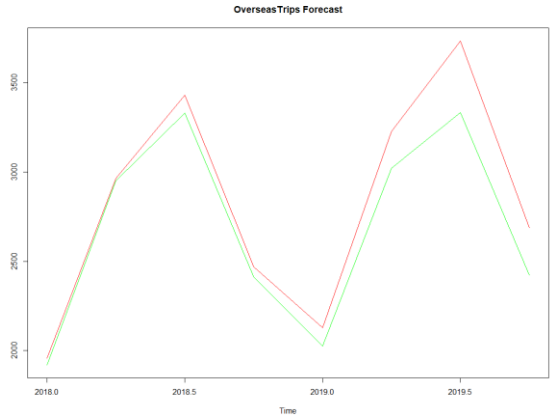

Fig. 7. Forecast Comparison

(Fig. 7) Shows the comparison of forecasts. We have used log10(test_data) because we had forecasted with decomposed data using STL log10(train_data).

The Green line shows the real observation and Red line shows the forecasted values.

B. Seasonal Naïve:

Next, Seasonal naïve has been used for forecasting.

```
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

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
2020 Q4         2424.6 2198.213 2650.987 2078.371 2770.829
2021 Q1         2026.7 1706.541 2346.859 1537.059 2516.341
2021 Q2         3021.8 2701.641 3341.959 2532.159 3511.441
2021 Q3         3334.4 3014.241 3654.559 2844.759 3824.041
2021 Q4         2424.6 2104.441 2744.759 1934.959 2914.241
```

Fig. 8. Summary of Seasonal Naive

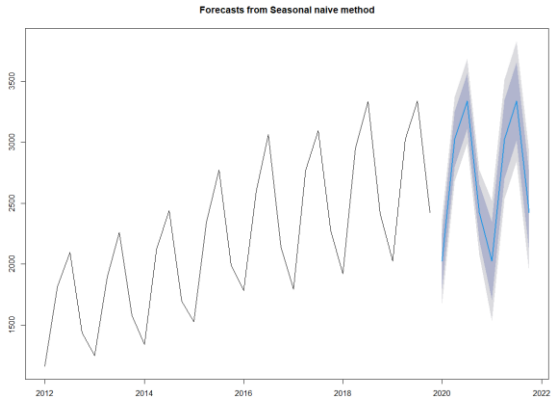

Fig. 9. Forecasts from Seasonal Naïve

(Fig. 8) Shows the Errors, residuals, and forecasted value. Here, RMSE is 176.65 and MAPE is 6.97. Lo/Hi 80 shows the 80 % confidence interval and Lo/Hi 95 shows 95 % confidence interval.

(Fig. 9) Shows the forecast using the Seasonal Naïve model. The blue line shows the average forecast, and the Grey part is Low and High region of the forecasts.

C. Holt-Winters:

Next, Holt-Winters has been used for forecasting. For using Holt-Winters, its function with model = "ZZZ" has been used. "ZZZ" model is used for automatic selection. Here, M,A,M is the best for this dataset. M is for Multiplicative and A is Additive.

(Fig. 10) Shows Smoothing parameters. Here, Alpha represents the level smoothing coefficient. Beta shows a trend smoothing coefficient. Gamma shows a seasonal smoothing coefficient. AIC, AICc, and BIC are also displayed. The value of RMSE is 54.69 and MAPE is 2.01.

```
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        ACF1
Training set -7.22914 54.69822 44.54334 -0.3013343 2.017642 0.2906987 -0.05187753
```

Fig. 10. Summary of Holt-Winters

(Fig. 11) Shows the forecast using Holt-Winters. The blue line is the average of the forecast and the grey area is the confidence interval of the forecast.

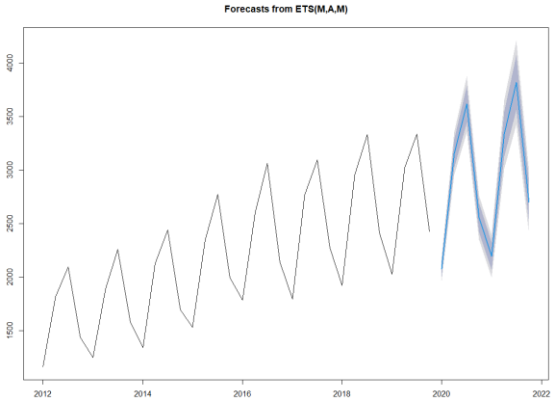

Fig. 11. Forecast from ETS(M,A,M)

Holt-Winters shows better forecast result than Seasonal Naïve.

D. ARIMA:

(Fig. 12), Shows ‘ggtsdisplay’ has been used to plot ACF and PACF plots.

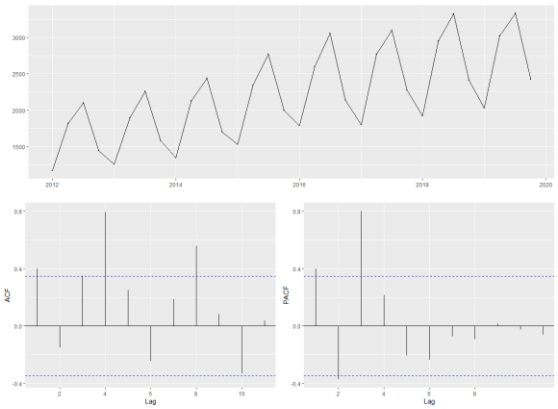

Fig. 12. Arima – ACF and PACF plots

ACF plot is an Auto Correlation Function that shows the correlation of time series with lagged values. ACF demonstrates how present value is related to past values of the series. ACF consider all components like trend, seasonality, residuals, etc. PACF is a Partial Auto Correlation Function, which demonstrates by showing the correlation of the residuals with the next lag value. ACF and PACF help in finding p,d,q values.

(Fig. 13) Differencing can be used to find the difference, which is 1. Differencing needs to be applied before using the ARIMA model.

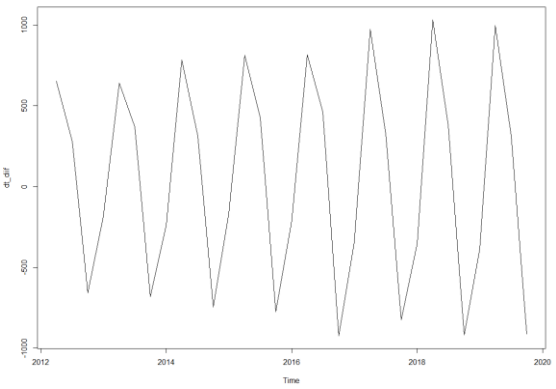

Fig. 13. Differencing plot against time

(Fig. 14) Shows KPSS Test for level stationarity. It is observed that KPSS Level = 0.083, Truncating lag parameter = 2, P-value = 0.1

```
        KPSS Test for Level Stationarity

data:  dt_diif
KPSS Level = 0.083767, Truncation lag parameter = 2, p-value = 0.1
```

Fig. 14. KPSS Test for Level Stationarity

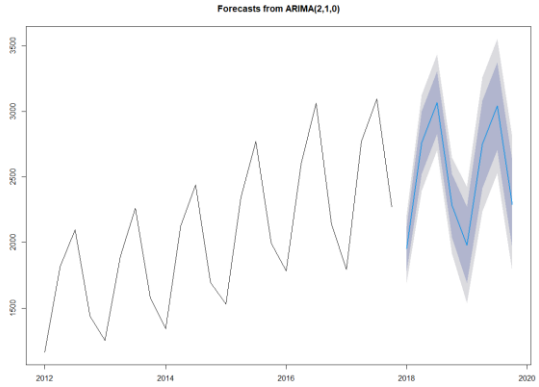

Fig. 15. Forecast from ARIMA (2,1,0)

(Fig. 15) Shows Forecast from ARIMA (2,1,0), where the blue line is average of forecast and grey out area is min and max range of the forecast.

```
Series: train_dt
ARIMA(2,1,0)

Coefficients:
          ar1      ar2
      -0.0013  -0.9758
s.e.   0.0466   0.0223

sigma^2 estimated as 17385:  log likelihood=-146.91
AIC=299.82   AICc=301.08   BIC=303.22

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 81.09398 123.3368 98.83093 3.899408 4.837519 0.5788218 -0.5118056
```

Fig. 16. Summary of ARIMA (2,1,0)

(Fig. 16) Shows the Summary of ARIMA (2,1,0) model, where RMSE is 123.33 and MAPE is 4.83.

From all the tests, we concluded that Holt-Winters shows the best result for this Seasonal “OverseasTrips” dataset.

B. NewHouseRegistrations_Ireland:

This is a dataset that contains information on the Registration of new houses in time 1978 to 2019.

(Fig. 17) Shows the time series plot. It is observed that trend and is not constant and there is a random change in the graph. Also, this is a non-seasonal dataset. It shows the Price of the new house went very high from the year 2002 to 2007.

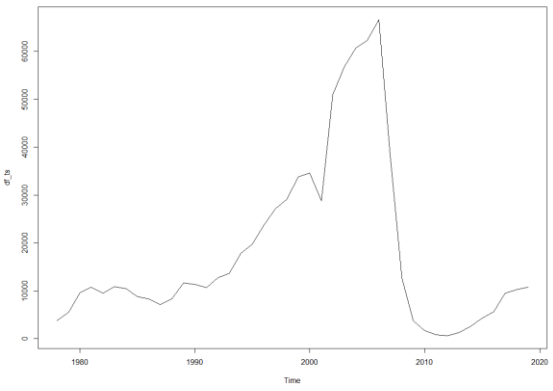

Fig. 17. Time Series Plot

A. SES (Seasonal Exponential Smoothing):

Next, we will use SES (Simple Exponential Smoothing) method because SES is best for forecasting a dataset that does not has any clear trend or seasonality. This data set has no clear trend and seasonality.

(Fig. 18) Shows Forecast from Seasonal Exponential Smoothing method.

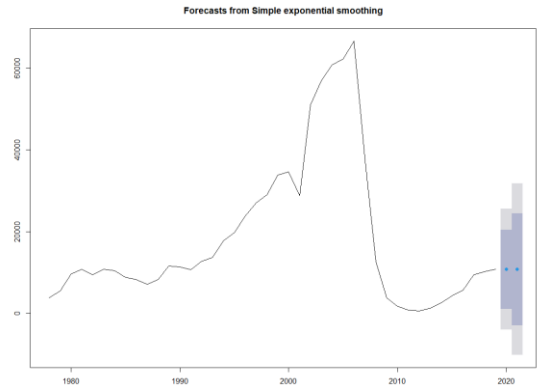

Fig. 18. Forecasts from SES

The Blue dotted line shows the average of forecasts and grey area is the range of its forecasts.

(Fig. 19) Shows the comparison of real data and forecasted data. The Red line shows forecasted data whereas black line shows the real observation.

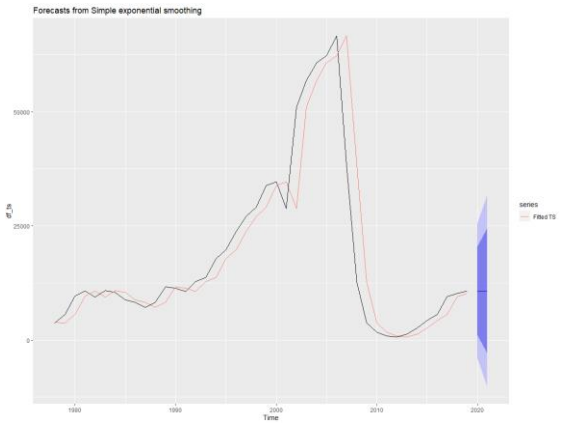

Fig. 19. Forecasts - SES compared with Actual data.

(Fig. 20) Shows summary of SES model. RMSE is 7378.82 and MAP is 33.33. We will look for another model because here we are getting high Errors. It is observed that only Alpha is present which shows the coefficient of level smoothing. Here, Beta and Gamma are missing because this dataset does not have Trend and Seasonality.

```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = df_ts, 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      ACF1
Training set 165.2083 7378.822 3875.789 -7.771767 33.33161 0.9769476 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. 20. Summary of SES Model

B. Holt with ETS:

Next, the Holt model is used for this dataset. First, need to create a subset of time series using the window() function. Here, the start value is 1978. Now, we will use this with ets() function to forecast, here, model = “ZZZ” and damped = “True”, the damped parameter will use either Multiplicative or Additive method.

We will create 2 models, one with damped True and one with False. Then we will compare the forecast to check the model.

(Fig. 21 and 22) Shows summary of ETS model with damped = TRUE and FALSE. It can be observed that for both the model, Alpha is 0.9999 which means the dataset has a high-level smoothing coefficient. The trend is missing and has a very small Beta value near 0.

```
ETS(M,Ad,N)

Call:
 ets(y = df_ts_holt, model = "ZZZ", damped = TRUE)

  Smoothing parameters:
    alpha = 0.9999
    beta  = 1e-04
    phi   = 0.98

  Initial states:
    l = 6960.8205
    b = 273.0356

  sigma:  0.3953

     AIC     AICc      BIC
868.9059 871.3059 879.3319

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -91.83774 7390.073 3902.742 -12.69181 35.90145 0.9837413 0.4165348
```

Fig. 21. Summary of ETS (damped = TRUE)

```
ETS(M,A,N)

Call:
 ets(y = df_ts_holt, model = "ZZZ", damped = FALSE)

  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

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set -184.0889 7395.984 3870.015 -14.83976 36.43003 0.975492 0.4173692
```

Fig. 22. Summary of ETS (damped = FALSE)

Both models have Errors near to each other. RMSE is 7390.07 and MAPE is 35.90 when damped is TRUE and ETS shows (M,ad,N) which means Errors are Multiplicative, Trend is damped and Season as None.

However, RMSE is 7395.98 and MAPE is 36.43 when damped is FALSE and ETS shows (M,A,N), which means Errors are multiplicative, Trend as additive and Season as None.

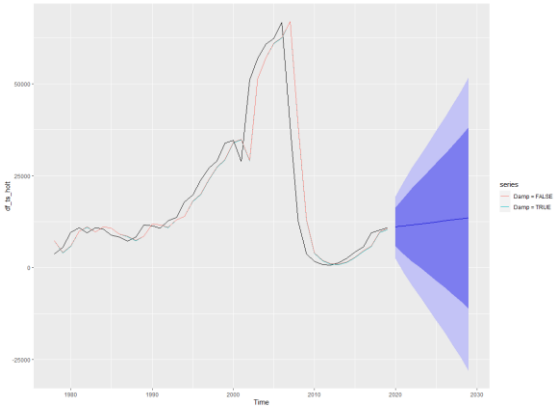

Fig. 23. Forecasts from ETS

(Fig. 23) Shows forecasts for both the Damped and non-damped model. The Red line shows forecast value when Damped = FALSE and the dark green line shows forecast value when Damped = TRUE. The Dark blue line shows an average of forecasts.

Next, we need to Stationaries time series. For this, SMA (Simple Moving Average) has been used with various n values like 1,3,5,7, and 10. (Fig. 24) Out of which, for n = 5, we got the best smoothing time series.

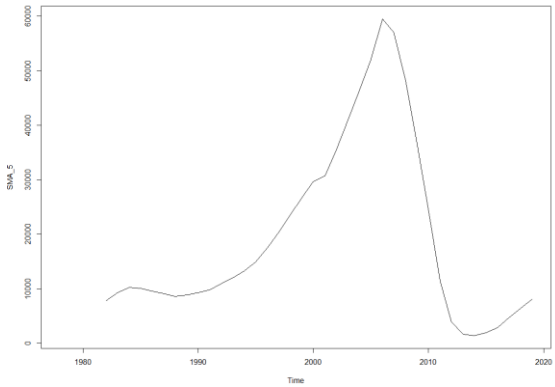

Fig. 24. Time Series Plot after SMA, n=5

Now, we used this SMA, n=5 time series with SES and ETS(Holt) models and found a better result than before. Errors for SES are RMSE: 4683.91 and MAPE: 24.66. Errors for ETS(Holt) are RMSE: 2540.424 and MAPE: 22.69.

C. AUTO ARIMA:

Next, we used the AUTO ARIMA model, for this, we are using a normal time series object, we are not using SMA, n = 5-time series. (Fig. 25) Shows the Summary of Auto ARIMA model.

```
Series: df_ts
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1      ar2       mean
      1.3346  -0.4665  16791.106
s.e.  0.1315   0.1319   6985.181

sigma^2 estimated as 43317727:  log likelihood=-428.43
AIC=864.86   AICc=865.94   BIC=871.81

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 207.1252 6342.208 3464.418 -20.20197 35.95662 0.8732557 -0.007018081
```

Fig. 25. Summary of AUTO ARIMA

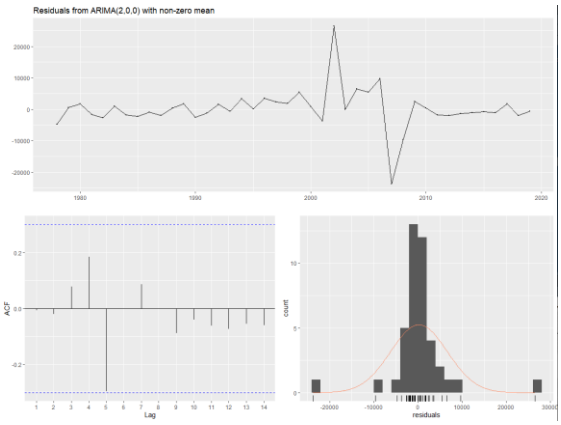

Fig. 26. Residuals from ARIMA (2,0,0) with non-zero mean

(Fig. 27) Shows the Normal Q-Q plot of our Auto ARIMA model, which shows residuals.

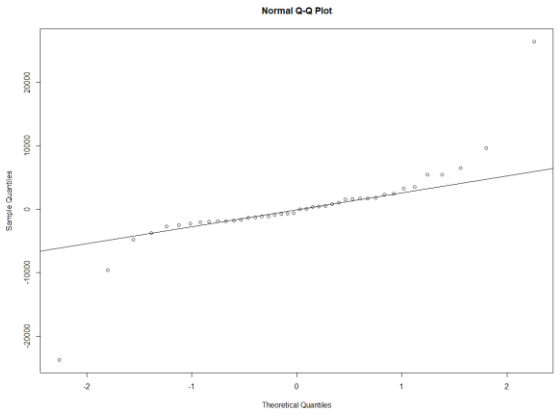

Fig. 27. Normal Q-Q Plot (AUTO ARIMA)

(Fig. 28 and 29) It Shows ACF and PACF plots. We got RMSE: 6342.20 and MAPE: 35.95 for Auto ARIMA model. (Fig. 30) Shows the forecast of Auto ARIMA.

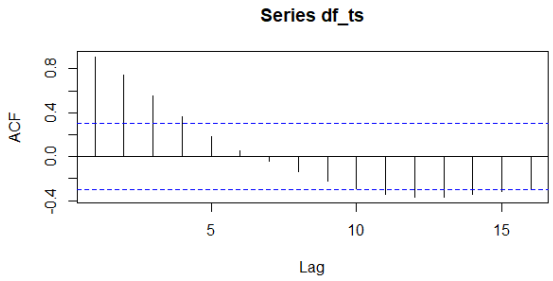

Fig. 28. ACF

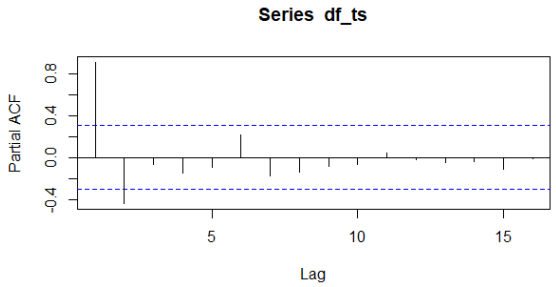

Fig. 29. PACF

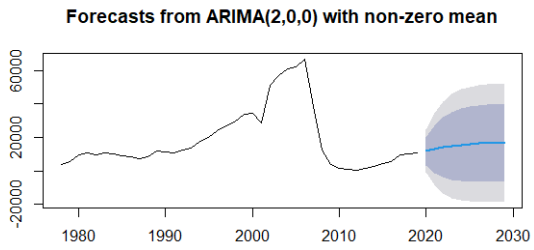

Fig. 30. Forecasts of AUTO ARIMA

When comparing all models, SES for both normal time series object and SMA, n =5 object, were best with results.

II. LOGISTIC REGRESSION

Logistic regression is a statistical model which can be used to create a model only if the dependent variable is binary (True/False, Yes/No, 1/0, etc.) Logistic Regression produces an S-shaped curve that takes values between 0 and 1. Using threshold value (i.e., 0.5 by default), values are filtered based on the threshold value and if the final values are less than 0.5, it will be assigned as 0 and if the final value is greater than 0.5, it will be assigned as 1. IBM SPSS has been used for this dataset.

This is a dataset of child weight. It has a total of 16 variables. We are removing 'ID' as it is useless. Also, after analysing data, 'lowbwt' is used as a Target Variable and the rest 14 as Independent Variables, however, based on Target Variable, it is clearly observed that 'Birthweight' is not useful for this Target Variable as it shows values which we want to predict. So, after removing 'Birthweight', we are processing our models by selecting the best variables out of 13 independent variables (Length, Headcirc, Gestation, mnocig, mage, mheight, mppwt, fage, fedyrs, fnocig, fheight, Smoker, mage35).

Steps used to build models:

1. A simple Logistic Regression model has been used for this dataset, where Target Variable is 'lowbwt' and the rest 13 variables are independent variables.

(Fig. 31) Shows Nagelkerke R Square value as 1. This value can be ranged from 0 to 1.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.560	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Fig 31. Model Summary

(Fig. 32) Shows Hosmer-Lemeshow goodness of fit test is used which shows P = 1.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	8	1.000

Fig. 32. Hosmer and Lemeshow Test

After observing the results (Fig. 33) of the Simple Logistic Model, it is observed that it is predicting 100 % Accuracy, 100 % Recall/Sensitivity and 100 % Specificity. It is clearly observed from the predicted values that our model is Over fitting. This can be possible because of small records in the dataset or wrong variable selection.

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. 33. Classification Table of Logistic Model

2. We must apply techniques to find the best variables out of 13. First, after observing the dataset, it is clearly visible that out of 13, 11 are numerical variables and 2 are categorical variables. For numerical variable selection, we are using the

ANOVA test and for categorical variable selection, we are using Chi-Square.

Applying ANOVA to 11 numerical variables (Length, Headcirc, Gestation, mnocig, mage, mheight, mppwt, fage, fedyrs, fnocig, fheight).

(Fig. 34) Shows ANOVA results, it can be clearly observed that P < 0.05 for Length, Headcirc, Gestation and mppwt, which rejects Null Hypothesis by accepting alternate Hypothesis H1.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Length	Between Groups	131.444	1	131.444	23.696	.000
	Within Groups	221.889	40	5.547		
	Total	353.333	41			
Headcirc	Between Groups	47.147	1	47.147	9.980	.003
	Within Groups	188.972	40	4.724		
	Total	236.119	41			
Gestation	Between Groups	104.143	1	104.143	22.847	.000
	Within Groups	182.333	40	4.558		
	Total	286.476	41			
mage	Between Groups	7.683	1	7.683	.235	.631
	Within Groups	1308.722	40	32.718		
	Total	1316.405	41			
mnocig	Between Groups	8.036	1	8.036	.050	.824
	Within Groups	6410.250	40	160.256		
	Total	6418.286	41			
mheight	Between Groups	68.099	1	68.099	1.635	.208
	Within Groups	1666.306	40	41.658		
	Total	1734.405	41			
mppwt	Between Groups	266.194	1	266.194	5.730	.021
	Within Groups	1858.306	40	46.458		
	Total	2124.500	41			
fage	Between Groups	116.036	1	116.036	2.556	.118
	Within Groups	1815.583	40	45.390		
	Total	1931.619	41			
fedyrs	Between Groups	7.000	1	7.000	1.519	.225
	Within Groups	184.333	40	4.608		
	Total	191.333	41			
fnocig	Between Groups	869.143	1	869.143	3.046	.089
	Within Groups	11413.333	40	285.333		
	Total	12282.476	41			
fheight	Between Groups	19.444	1	19.444	.393	.534
	Within Groups	1977.056	40	49.426		
	Total	1996.500	41			

Fig 34. ANOVA

Applying Chi-Square on categorical variables, i.e., Smoker and mage35.

(Fig. 35a & 35b) Shows Chi-Square results, it can be clearly observed that both Smoker and mage35 have P > 0.05 which means, it fails to reject the Null Hypothesis H0.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.689[a]	1	.101		
Continuity Correction[b]	1.436	1	.231		
Likelihood Ratio	2.927	1	.087		
Fisher's Exact Test				.187	.115
Linear-by-Linear Association	2.625	1	.105		
N of Valid Cases	42				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.86.

b. Computed only for a 2x2 table

Fig. 35a. Chi Square test for Smoker * lowbwt

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. 35b. Chi Square test for mage35 * lowbwt

Based on the above ANOVA and Chi-Square tests, it can be clearly observed that Length, Headcirc, Gestation and mppwt are the best variables for predicting Target Variable 'lowbwt'. Rest all variables had P > 0.05 which means it fails to reject Null Hypothesis H0, which eventually means that there are some relations between variables, which can impact in predicting the dependent variable.

Now, after performing the above ANOVA and Chi-Square, only 4 variables can be used, i.e., Length, Headcirc, Gestation and mppwt. We will use only 4 variables for Logistic regression.

(Fig. 36) Shows the result of this model with default classification cut-off (threshold) 0.5.

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6

a. The cut value is .500

Fig. 36. Classification Table

After, applying different classification cut-off values like 0.3, 0.4, 0.6, and 0.7, it can be clearly demonstrated that the Sensitivity/Recall of all models is 83.3 % whereas there is a change in Accuracy and Specificity. Below (Fig. 37) shows a table that contains results based on all threshold values.

Cut-off	Accuracy	Recall	Specificity
0.3	95.2	83.3	97.2
0.4	95.2	83.3	97.2
0.5	97.6	83.3	100
0.6	97.6	83.3	100
0.7	97.6	83.3	100

Fig. 37. Results of Logistic Regression for selected variables

The above results demonstrate that the Model is best in predicting values when the Cut-off is 0.5, 0.6 or 0.7. Here, we are getting the highest accuracy, true

positive rate (Recall) and true negative rate (Specificity).

3. Next, PCA (Principal Component Analysis) is being used for dimension reduction. PCA must be used in 2 ways:

3A. PCA will be applied to only 4 variables, i.e., Length, Headcirc, Gestation and mppwt because, in the above step, we used ANOVA and Chi-Square to select the best variables. KMO and Barlett tests are used. If the KMO value is less than 0.50, the result is probably useless. If Barlett's P > 0.05, it indicates that factor analysis may be good to use. (Fig. 38.) Shows KMO and Barlett tests result, it can be observed that both KMO and Barlett's factor analysis is not useful for our data.

KMO and Bartlett's Test[a]

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.407
Bartlett's Test of Sphericity	Approx. Chi-Square	4.276
	df	6
	Sig.	.639

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

Fig. 38. KMO and Barlett's Test

(Fig. 39) Shows total variance explained. It can observe that 4 components are used and only 1st component has an Eigen value greater than 1. This 1st component is having 54.8 % initial variance and a cumulative of 54.8 %.

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.194	54.853	54.853	2.194	54.853	54.853
2	.971	24.286	79.139			
3	.682	17.061	96.200			
4	.152	3.800	100.000			

Extraction Method: Principal Component Analysis.

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

Fig. 39. Total Variance Explained

(Fig. 40) Shows the Scree plot, which shows all 4 components against Eigen Values.

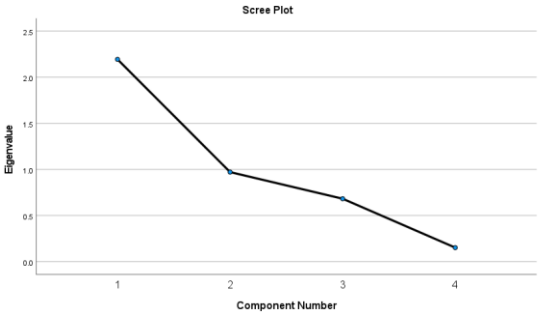

Fig. 40. Scree Plot

However, after using Logistic Regression with different classification cut-off (Threshold), it can be observed that the model is predicting good results. The model shows the same results as the before model for the best cut-off of 0.5 and 0.6. (Fig. 41) Shows result when the cut-off is 0.5.

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6

a. The cut value is .500

Fig. 41. Classification Table

(Fig. 42) Shows results for classification cut-off values 0.3, 0.4, 0.5, 0.6, and 0.7.

Cut-off	Accuracy	Recall	Specificity
0.3	90.5	83.3	91.7
0.4	95.2	83.3	97.2
0.5	97.6	83.3	100
0.6	97.6	83.3	100
0.7	95.2	66.7	100

Fig 42. Results of Logistic Regression

The above fig. demonstrates that for cut-off value 0.5 and 0.6, the model is predicting best.

However, it may be good or may not be good because we had high values of KMO and Barlett's Test. But after comparing results with the previous Logistic Regression, we are getting the same result for cut-off value 0.5 and 0.6.

3B. PCA will be applied to all 11 Numerical Variables. We are not using Smoke and Mage35 as those are categorical variables and during the Chi-Square test, they had P > 0.05, which means it fails to reject Null Hypothesis H0.

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.584	41.677	41.677	4.584	41.677	41.677
2	2.868	26.071	67.748	2.868	26.071	67.748
3	2.286	20.781	88.529	2.286	20.781	88.529
4	1.049	9.536	98.065	1.049	9.536	98.065
5	.213	1.935	100.000			
6	1.259E-15	1.145E-14	100.000			
7	4.947E-16	4.497E-15	100.000			
8	1.692E-16	1.538E-15	100.000			
9	9.341E-17	8.492E-16	100.000			
10	-5.111E-17	-4.646E-16	100.000			
11	-3.419E-16	-3.108E-15	100.000			

Extraction Method: Principal Component Analysis.

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

Fig. 43. Total Variance Explained

(Fig. 43) shows that out of 11 components, only 4 components are having Eigenvalues greater than 1.

(Fig. 44) Shows the Scree plot and it can be easily observed that the first 4 variables have an Eigen-value > 1 and are above the elbow.

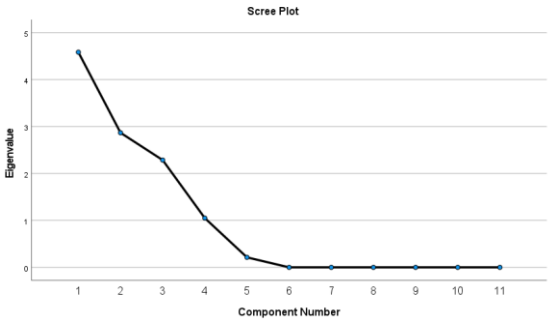

Fig. 44. Scree Plot

Again, the Logistic model is used with different values of classification cut-off like 0.3, 04, 0.5, 0.6, and 0.7.

(Fig. 45) Shows classification table for default cut-off value 0.5.

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	35	1	97.2
		1	2	4	66.7
	Overall Percentage				92.9

a. The cut value is .500

Fig. 45. Classification Table

Below (Fig. 46) Shows results for all applied cut-off values.

Cut-off	Accuracy	Recall	Specificity
0.3	90.5	66.7	94.4
0.4	92.9	66.7	97.2
0.5	92.9	66.7	97.2
0.6	95.2	66.7	100
0.7	95.2	66.7	100

Fig 46. Results of Logistic Regression

From the above (Fig. 46), it can be observed that the model is predicting good for cut-off value 0.6 and 0.7. Here, the model is showing accuracy as 95.2 %, Sensitivity/Recall (True positive rate) as 66.7 % and Specificity (True negative rate) as 100 %. That means the model is predicting good for babies who are not actually low weight, and the model is predicting 100 % of the times if it's underweight or not. However, recall is only 66.7 %, which means the model is predicting bad when babies are actually having ow weight and the model fail to predict all and only have a positive rate of 66.7 %. However, with this dataset, we are getting the best accuracy of 95.2 %.

CONCLUSIONS

Part A: We used 2 time-series datasets.

First, OverseasTripe was a Seasonal Time Series data. Few models used like Basic Forecast, Seasonal Naïve, Holt-Winters, and ARIMA to find the best model for this dataset. Out of all, Holt-Winters showed the best result with a good forecast and fewer errors, where RMSE is 54.69 and MAPE is 2.01.

Second, NewHouseRegistrations_Ireland, was a non-seasonal dataset. Few models used like SES, ETS (Holt), and SMA. n=5 for both SES and ETS(Holt) again and in last used Auto ARIMA. ETS(Holt) showed the best result for both normal time series data and SMA, n=5 smoothing data, Errors for ETS(Holt) are RMSE: 2540.424 and MAPE: 22.69.

Part B – Child Births dataset. First, we used simple logistic regression with all independent variables, however, we got 100% accuracy, which was because of Overfitting. Next, Variables are selected with help of ANOVA, Chi Sq. After selecting few variables, Logistic Regression was applied with various classification cut-off like 0.3,0.4,0.5,0.6, and 0.7. Logistic Model showed good result when variables are selected and cut-off were 0.5,0.6, and 0.7. Accuracy: 97.6 %, Recall: 83.3 %, Specificity: 100 %. Also, PCA has been used. However, after applying dimension reduction PCA to 11 numerical variables, it was observed that the model predicted good when the categorical cut-off is 0.6 and 0.7, where Accuracy: 95.2 %, Recall: 66.7 % and Specificity: 100 %.

REFERENCES

[1] Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models. IMECS 2013, March 13 - 15, 2013, Hong Kong

[2] Logistic Regression, Ch 5, Daniel Jurafsky & James H. Martin. December 30, 2020. stanford.edu