Statistical Analysis Of Time Series and Logistic Regression

Aryan Rajput
(Data Analytics)
National College of Ireland
Dublin, Ireland
x20128088@student.ncirl.ie

***Abstract*—The terminal assignment is divided in two sub-parts which include Time Series analysis and Logistic Regression. Time series analysis has been performed upon two given datasets, which are OverseasTrips and NewHouseRegistrations. In this analysis the components of time series are evaluated and then three models have been applied on each of the dataset. Both the time series predict the data for 3 future periods. Then all the models have been compared with each other on certain parameters to get the best fit model. Whereas Logistic Regression has been performed upon given dataset ChildBirth. Smoker has been kept as independent variable and Principal Component Analysis (PCA) has been used to reduce dimensions during the process. Time Series analysis has been performed using R language in R studio whereas Regression task has been done in Statistical Package for the Social Sciences(SPSS)**

***K*eywords – Time series, Logistic Regression, ARIMA, Naive Model**

I. Introduction

In this terminal assignment, time series and logistic regression are parts of our statistical analysis. In time series analysis two datasets have been provided of New House Registration in Ireland which includes annual data from a period of 1978-2019. For analysis of this data we have performed Naive, Simple Exponential Smoothing Model and ARIMA model. The models have been compared and analysed on the basis of Root Mean Square Error(RMSE).The other given dataset consist quarterly data of overseas trips from year 2012 till 2019. For this data we have performed seasonal Naive, SARIMA and ETS models. These models were evaluated on basis of RMSE, AIC value and Ljung-Box test value. Logistic regression is performed and three models have been prepared using different combination of variables. Principal Component Analysis has been performed for dimension reduction and again model has been made.

II. Time Series Analysis

Time series analysis has been done over both the datasets. The analysis has been discussed below one by one.

A. *OverseasTrips*

Overseas tips dataset consist of number of trips in thousands recorded quarterly starting from year 2012.

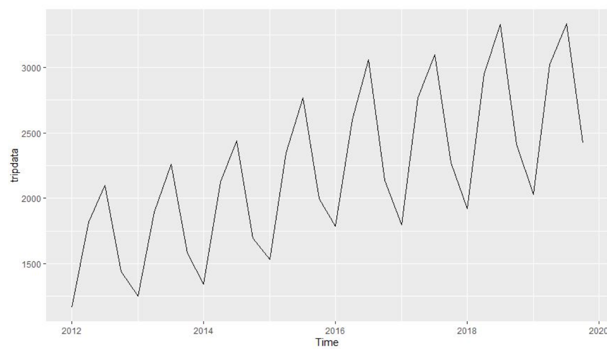

Fig. 1. Plot of Dataset for overseas trips

1) Data Analysis: A basic plot has been made to understand the data and to get insights from data.

As seen in the figure there is a inclined trend in the data as well as cyclic patterns are also observed in the data thus it can be said that data is seasonal. Moreover, as data seems to be increasing over time, so the given data is said to be multiplicative.

To confirm the seasonality of the data, seasonplot has been plotted to understand data more vividly.

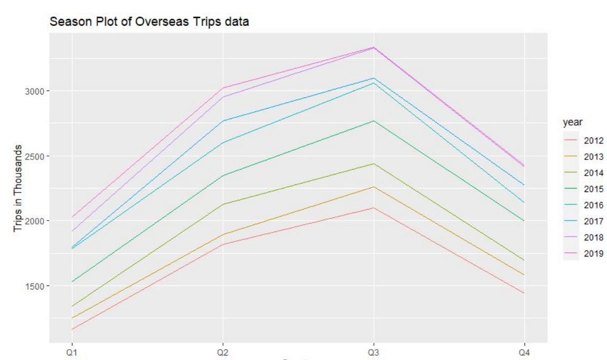

Fig. 2. Seasonal Plot for overseas trips

As clearly seen in the Figure 2, data seems to be increasing from quarter 1 till quarter 3 but decreasing significantly for quarter 4. Thus it can be established that the data is seasonal.

The data is multiplicative, hence decomposition of data has been done by taking seasonal multiplicity into account.

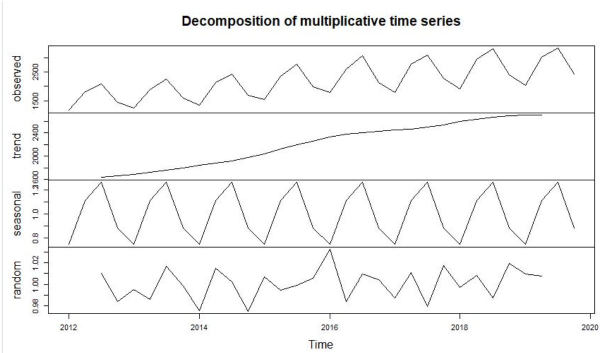

Fig. 3. Decomposition Plot for overseas trips

It can be observed from the decomposition plot that the given data contains a linear trend as well as seasonality.

2) Model Fitting: Four models have been fitted on the data as discussed in details below:

- **Seasonal Naive Model:** One of the simple model is Seasonal Naive model which is used for highly seasonal data and forecast values as the last observed value from the same season of the year.
  For our data, Seasonal Naive model predicted with a root mean square error (RMSE) value as 176.65

```
Forecast method: Seasonal naive method

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

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
```

Fig. 4. Seasonal Naive model Output for overseas trips data

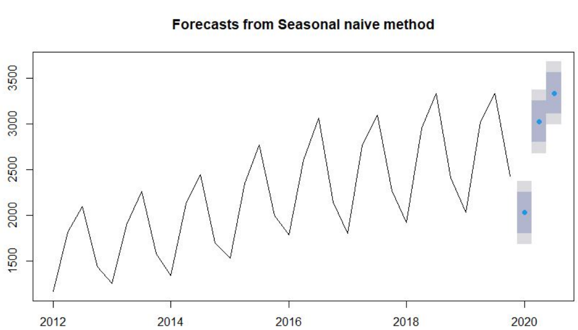

Fig. 5. Seasonal Naive model plot for Trips data

- **Seasonal ARIMA Model:** Method ndiffs has been used to determine the number of differences required for time series to be made stationary. For this series it returned output as 1. Also for seasonal differences nsdiffs function is used and it also returned value as 1 which is taken as D=1 while applying SARIMA. To verify whether our data is stationary or not, Augmented Dickey-Fuller (ADF) Test has been performed, but it returned p-value as 0.99 which resulted to failed our null hypothesis that data has unit root. Thus it can be said that the data is trending. It is handled later while appliying ARIMA model.

  ARIMA is abbreviation for 'Auto Regressive Integrated Moving Average', which is used to explain and forecast data based upon its past values and lagged forecast errors. SARIMA is extension of ARIMA model, which handles the direct modeling of the seasonal component. It is characterized by terms SARIMA(p,d,q)(P,D,Q)m. Where p can be described as the order of the AutoRegression, q is term of moving averages, d is the defferencing required to make the time series stationary and P,D,Q are related to the similar seasonal factor of all three mentioned as before. The m factor denotes number of steps for a single seasonal cycle.
  After applying auto ARIMA the results obtained are displayed in Fig 6

```
ARIMA(1,0,0)(0,1,0)[4] with drift

Coefficients:
         ar1    drift
      0.5835  35.9414
s.e.  0.1585   7.9346

sigma^2 estimated as 5616:  log likelihood=-159.77
AIC=325.53   AICc=326.53   BIC=329.53
```

Fig. 6. SARIMA model Output for Overseas Trips data

  Auto ARIMA applied by taking p=1 and D=1 and as the data is quarterly m = 4 is taken into account.
  The prediction has been made with an RMSE value of 67.54 and Akaike Information Criterion (AIC) value as 325.53
  The prediction for three consecutive quaters using SARIMA model is displayed in Fig 7.

```
AIC=325.53   AICc=326.53   BIC=329.53
> tripdata%>%auto.arima()%>%accuracy
                   ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set 1.570482 67.54754 55.32141 -0.09590574 2.451533 0.3610385 -0.0396672
> fcast <- forecast(aamodel, h=3)
> fcast
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2093.465 1997.429 2189.501 1946.590 2240.339
2020 Q2       3120.636 3009.447 3231.825 2950.587 3290.685
2020 Q3       3451.950 3336.053 3567.846 3274.701 3629.198
```

Fig. 7. Prediction using SARIMA model

  The plot for the predictions is displayed in Fig 8

  After the prediction the residuals plot has been made to analyse the difference among regression line vertically

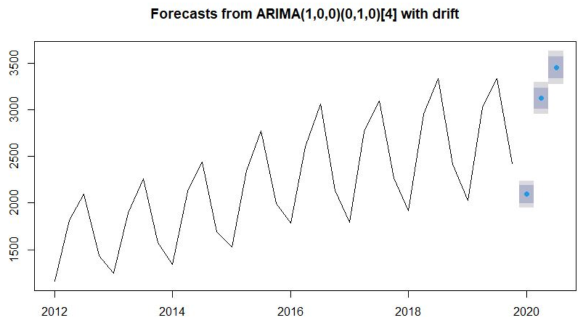

Fig. 8. Prediction using SARIMA model

missing a data point. And the residual values should randomly and equally spaced around the horizontal axis. By plotting the residual plot, the residuals has been verified and noted that all the residuals seems to be spread all over significantly. From residual ACF plot it can also be interpreted that all the residuals falls under limit and distributed significantly.

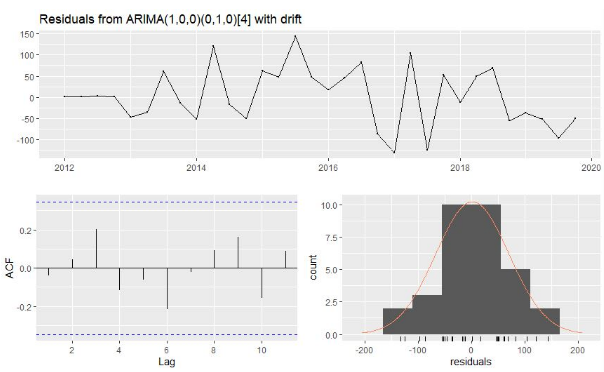

Fig. 9. Residual plot for SARIMA model

The QQ plot has been made to check the scattered quantile values of the distribution.

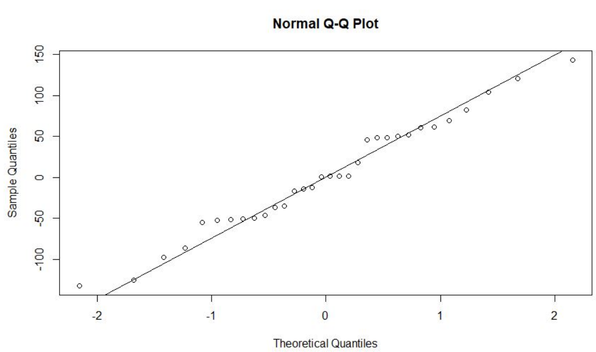

Fig. 10. Q-Q plot for SARIMA model

As seen in Fig.11 all the points are plotted nearly to the mean line and distributed equally.

```
        Box-Ljung test

data:  aamodel$residuals
X-squared = 0.055224, df = 1, p-value = 0.8142
```

Fig. 11. L-Jung box Test plot for SARIMA model

L-Jung Box Test has been performed to check the dependency of residuals. In this test the p-value is observed as 0.8142, which is quite significant and considered as good for the test.

- **ETS Model:** ETS is an exponential model where ets is abbreviation used for error, trend and seasonal model. This model is very powerful exponential model and uses 3 letters to model. The available options are as listed below:
  A - additive
  M - multiplicative
  N - none
  Z - automatically
  Any possible combination could be used for the model. As our data has multiplicative seasonality thus, firstly MNM model has been implemented.

```
  sigma:  0.0478

     AIC     AICc      BIC 
413.2172 417.8839 423.4774 

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 28.84382 78.43142 60.80743 1.155509 3.017543 0.3968413 -0.1912264
```

Fig. 12. ETS(MNM) Output for House Reg. data

For our data, ETS(MNM) predicted with a root mean square error (RMSE) value as 78.43 and AIC value as 413.21

- **ETS auto Model:** To check the best fit model zzz has been used to fit the model and the output obtained from this model is displayed in Fig 13.

```
ETS(M,A,M) 

Call:
 ets(y = tripdata, 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. 13. ETS(MAM) Output for House Reg. data

For our data, automatically used values for ETS are MAM which predicted values with a root mean square error (RMSE) value as 54.698 and AIC value as 378.82

B. New House registration

New House registration data consist of number of annually registered houses over a 42 years of period starting from 1978.

1) Initial Data Analysis: Different checks have been performed on the dataset to check for seasonality and trend in the data. Data has been smoothed using Moving averages for 3 and 5 level of k, to analyse the data in better manner.

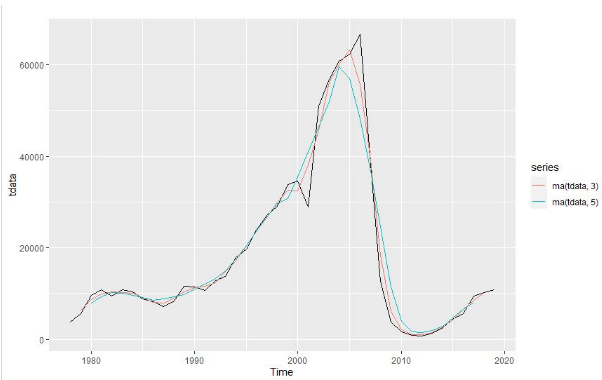

Fig. 14. Plot of Dataset after Moving Averages

As seen in the figure there is a trend for early phase in the data but near year 2007 data drop is breaking the trend. No cyclic patterns have been observed so it can be defined there is no seasonality in the data. At k value of 5 for moving average data seem to be a little over-smoothened so moving average at 3 level has been taken into consideration for analysing the data.

Method ndiffs has been used to determine the number of differences required for time series to be made stationary. For this series it returned zero so no differences was needed for this data. To verify whether our data is stationary or not, Augmented Dickey-Fuller (ADF) Test has been performed, but it returned p-value as 0.5243 which resulted to failed our null hypothesis that data has unit root. Thus it can be said that the data has local trend. To confirm the data stationarity, KPSS test has been performed which returned p-value as 0.1, so it is derived that data is stationary.

2) Model Fitting:

- **Naive Model:** One of the simple model is Naive model which forecast values as the last observed value. For our data, Naive model predicted with a root mean square error (RMSE) value as 7466.73 .

```
Error measures:
                   ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 170.8049 7466.737 3967.244 -7.900382 34.07645    1 0.4216754

Forecasts:
     Point Forecast     Lo 80    Hi 80      Lo 95    Hi 95
2020          10784  1214.992 20353.01  -3850.535 25418.54
2021          10784 -2748.621 24316.62  -9912.358 31480.36
2022          10784 -5790.009 27358.01 -14563.759 36131.76
```

Fig. 15. Naive model Output for House Reg. data

As seen from the summary output the prediction for all the consecutive 3 years is 10784, the more detailed results and comparison has been made later in the report.

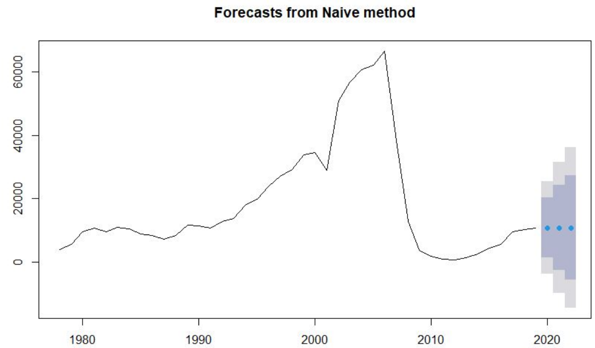

Fig. 16. Naive model plot for House Reg. data

- **Simple Exponential Smoothing Model:** SES model is used to forecast univariate data without trend or seasonality. SES requires a smoothing factor/coefficient. In this data alpha value is taken as 0.9995 and it predicted values with an RMSE of 7378.82.

```
  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
2022       10783.75 -5994.189 27561.69 -14875.894 36443.39
```

Fig. 17. SES model Output for House Reg. data

From the summary output the prediction for all the consecutive 3 years is 10783.75

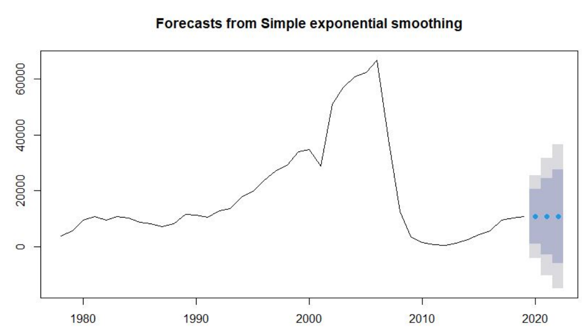

Fig. 18. SES model plot for House Reg. data

- **ARIMA Model:** ARIMA is abbreviation for 'Auto Regressive Integrated Moving Average', which is used to explain and forecast data based upon its past values and lagged forecast errors. It is characterized by three terms p,d,q where p can be described as the order of the AutoRegression, q is term of moving averages and d is the defferencing required to make the time series stationary. The correlation between observations of the time series has been measured by function ACF, whereas partial autocorrelation function (PACF) is use to get correlation with its own lagged values. The ACF and PACF values decide what values should be taken for auto regression and moving average in ARIMA model.

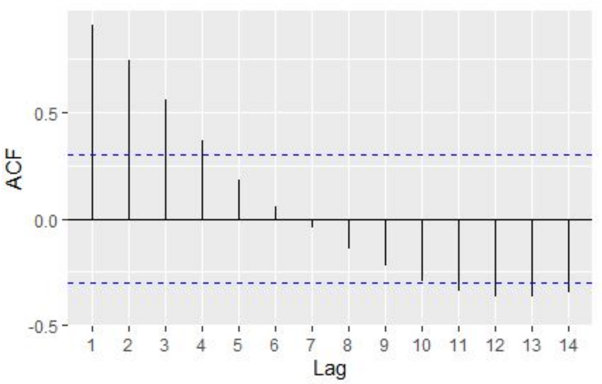

Fig. 19. ACF plot for House data

As we traverse through x-axis on ACF plot data seems to be decreasing continuously and making sinusoidal wave pattern. Though exceeding values for 4 logs as trend seems to be decreasing we are taking our q value as 0.

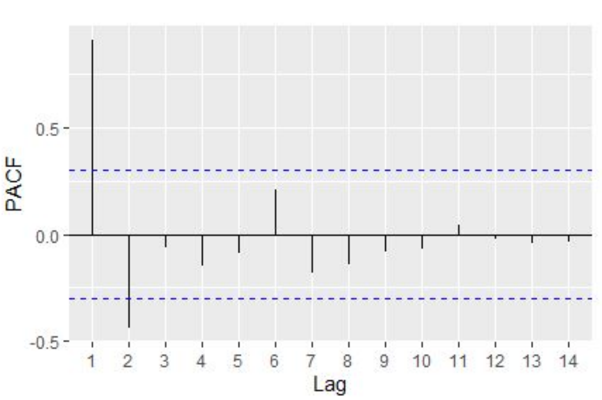

Fig. 20. PACF plot for House data

Whereas, for the PACF plot only for lag 1 and lag 2 data seems to be exceeding values significantly, thus giving our q value as 2.

As already checked above in data analyses using ndiff function we get differencing required as 0. So overall we get all required values of p,d,q for fitting ARIMA model as (2,0,0).

After applying ARIMA(2,0,0) the predicted results obtained are displayed in Figure 21

```
Series: tdata
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. 21. ARIMA model Output for House Reg. data

The prediction for future 3 years has been made with an RMSE value of 6342.20

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2020	11818.59	3383.902	20253.27	-1081.151	24718.33
2021	12957.23	-1109.161	27023.62	-8555.457	34469.91
2022	13994.20	-3917.264	31905.66	-13399.019	41387.42

Fig. 22. Prediction using ARIMA model for House Reg. data

The residuals plot has been made to analyse the difference among regression line vertically missing a data point. And the residual values should randomly and equally spaced around the horizontal axis. By plotting the residual plot, the residuals has been verified and noted that all the residuals seems to be spread over significantly.

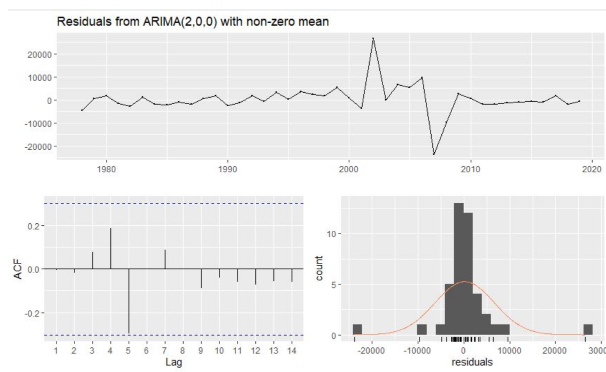

Fig. 23. Residual plot for ARIMA model

The QQ plot has been made to check the scattered quantile values of the distribution.

As seen in Fig. 24 all the points are plotted nearly to the mean line and distributed equally.

L-Jung Box Test has been performed to check the dependency of residuals. In this test the p-value is observed as 0.9624.

- **Auto ARIMA Model:** Auto ARIMA model has been made to confirm the input values of p,d and q that were provided to ARIMA model. Which resulted in the same output that we provided.

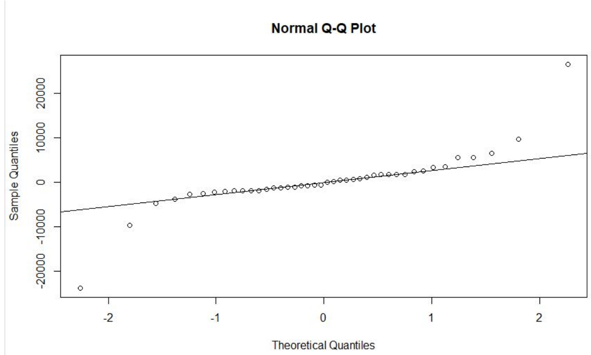

Fig. 24. Q-Q plot for ARIMA model

```
        Box-Ljung test

data:  fit$residuals
X-squared = 0.00222, df = 1, p-value = 0.9624
```

Fig. 25. L-Jung Box Test Result

```
Series: tdata
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
> tdata%>%auto.arima()%>%accuracy
                   ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 207.1252 6342.208 3464.418 -20.20197 35.95662 0.8732557 -0.007018081
```

Fig. 26. Auto ARIMA output

III. Time Series Results and Model Comparison

The comparison among Overseas Trips is done on basis of RMSE and AIC values. As for the models which were from same class their AIC value has been taken into account where as the models belonging to the different class are compared by only RMSE value.

Model Name	RMSE Value	AIC Value
Seasonal Naïve Model	176.6505	
SARIMA Model	67.54754	
ETS-MNM	78.43142	413.2172
ETS-MAM	54.69822	378.8267

Fig. 27. All Test Result Comparison for OverseasTrips

By comparing the RMSE value from all the models it can be observed that RMSE value of ETS(MAM) at 54.69 is least among all four models. Also it has lower AIC value(378.8267) as compared with ETS(MNM) and thus it can be established that from these 4 models ETS(MAM) model performed the best and could be considered as optimum model for overseas trip time series.

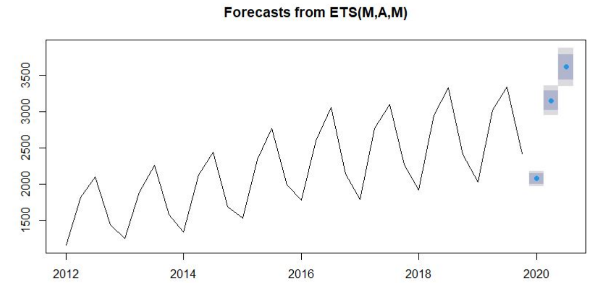

Fig. 28. Prediction of 3 quarter for OverseasTrips using ETS(MAM)

The ETS(MAM) model predicted values as 2076.043 for 2020 Q1, 3154.360 for 2020 Q2 and 3614.582 for 2020 Q3.

The comparison among New House Registration is done on basis of RMSE only because all the models belong to different class of the models.

Model Name	RMSE Value
Naïve Model	7466.737
SES Model	7378.822
ARIMA Model	6342.208
Auto ARIMA	6342.208

Fig. 29. All Test Result for HouseRegistration

As ARIMA model that we applied manually is same as the auto ARIMA, so among Naive, SES and ARIMA model, the RMSE value ARIMA is minimum. So it can be said that ARIMA is most optimum model for new house registration time series.

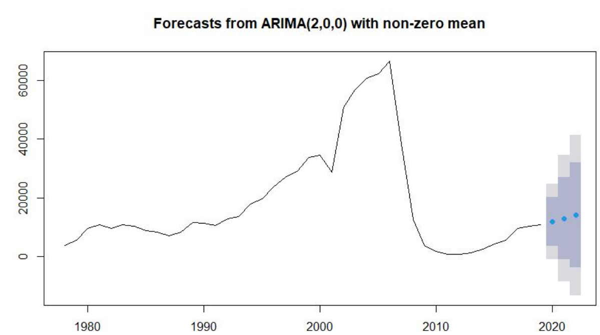

Fig. 30. Prediction from ARIMA model for House Reg. data

ARIMA model predicted values for consecutive years as 11818.59 in 2020, 12957.23 in 2021 and 13994.20 in 2022

IV. Logistic Regression

Logistic Regression is a predictive analysis and it is applied appropriately when the dependent variable is dichotomous (binary). This regression provides probability for a binary question (Yes vs No). The curve that this analysis has is S-shaped curve which takes any input and turn into a value between 0 and 1.

The analysis is performed upon childbirth dataset which has data related to childbirth and physical features of child, mother's smoking data ie. mother is smoker or not, number of mother's cigarette also father's details like age, education years etc.

For logistic regression mother smokes is taken as dependent variable which is also binary in nature. SPSS is used for performing logistic regression on this data.

A. Model-1

For first model building all the independent variables are taken into account. By applying logistic regression on this data resulted in below outcomes:

Classification Table[a,b]

	Observed		Predicted: smoker 0	Predicted: smoker 1	Percentage Correct
Step 0	smoker	0	0	20	.0
		1	0	22	100.0
	Overall Percentage				52.4

Fig. 31. Classification Table

Above classification table is from Block 0 and gives basic idea how distributed the given data is.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	38.818[a]	.369	.492

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

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.177	8	.627

Fig. 32. Model Summary Hosmer, Lemeshow Test

R2 value of the model fitted is .49 and Hosmer, Lemeshow Test has been taken into account for our test to check for the goodness of the fitted model. In our analysis we received value as .627

Classification Table[a]

	Observed		Predicted: smoker 0	Predicted: smoker 1	Percentage Correct
Step 1	smoker	0	16	4	80.0
		1	6	16	72.7
	Overall Percentage				76.2

a. The cut value is .500

Fig. 33. Classification Table after Model Fitting

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Length	-.109	.240	.208	1	.648	.896	.560	1.434
	Birthweight	-3.019	1.627	3.442	1	.064	.049	.002	1.186
	Headcirc	-.019	.256	.006	1	.939	.981	.594	1.618
	Gestation	.455	.273	2.771	1	.096	1.575	.923	2.690
	mage	-.012	.153	.006	1	.938	.988	.732	1.334
	mheight	.144	.110	1.699	1	.192	1.155	.930	1.433
	mppwt	-.030	.095	.098	1	.754	.971	.805	1.170
	fage	.130	.128	1.029	1	.310	1.139	.886	1.464
	fedyrs	.025	.228	.012	1	.912	1.025	.655	1.604
	fnocig	.072	.034	4.523	1	.033	1.074	1.006	1.147
	fheight	-.041	.082	.243	1	.622	.960	.817	1.128
	lowbwt	1.233	2.232	.305	1	.581	3.432	.043	272.617
	Constant	-21.063	16.957	1.543	1	.214	.000		

a. Variable(s) entered on step 1: Length, Birthweight, Headcirc, Gestation, mage, mheight, mppwt, fage, fedyrs, fnocig, fheight, lowbwt.

Fig. 34. Variable Summary

After model fitting overall achievement of the model is 76.2

By taking figure 34 into consideration it could be said that 11 out of 13 variables are highly significant.

B. Model-2

As Model 1 consisted all the independent variables, 11 variables has a high significance value. So by this it was concluded that the model is over fitted and to get rid of this problem, Principle Component Analysis(PCA) has been performed for dimension reduction.

To apply PCA several assumptions has to be satisfied. First assumption is that there should be multiple continuous variables in the data which were already present in our data. Also linear relationship should be there among variables which is confirmed by correlation matrix Fig 35.

Correlation Matrix

		Length	Birthweight	Headcirc	Gestation	mage	mnocig	mheight	mppwt	fage	fedyrs	fnocig	fheight	lowbwt	mage35
Correlation	Length	1.000	.727	.563	.705	.075	-.040	.485	.398	.137	.079	.009	.208	-.610	.131
	Birthweight	.727	1.000	.685	.708	.000	-.152	.363	.401	.176	.071	-.093	.031	-.652	-.109
	Headcirc	.563	.685	1.000	.405	.146	-.133	.337	.303	.301	.124	-.047	.042	-.447	.055
	Gestation	.705	.708	.405	1.000	.011	.043	.211	.255	.142	.131	-.114	.208	-.603	.007
	mage	.075	.000	.146	.011	1.000	.340	.060	.274	.807	.442	.091	-.200	-.076	.693
	mnocig	-.040	-.152	-.133	.043	.340	1.000	.126	.149	.248	.199	.257	.021	.035	.291
	mheight	.485	.363	.337	.211	.060	.126	1.000	.681	-.080	.035	.048	.274	-.198	.116
	mppwt	.398	.401	.303	.255	.274	.149	.681	1.000	.256	.180	.057	.093	-.354	.137
	fage	.137	.176	.301	.142	.807	.248	-.080	.256	1.000	.300	.136	-.269	-.245	.351
	fedyrs	.079	.071	.124	.131	.442	.199	.035	.180	.300	1.000	-.263	.018	-.191	.279
	fnocig	.009	-.093	-.047	-.114	.091	.257	.048	.057	.136	-.263	1.000	.329	.266	-.089
	fheight	.208	.031	.042	.208	-.200	.021	.274	.093	-.269	.018	.329	1.000	.099	-.188
	lowbwt	-.610	-.652	-.447	-.603	-.076	.035	-.198	-.354	-.245	-.191	.266	.099	1.000	.099
	mage35	.131	-.109	.055	.007	.693	.291	.116	.137	.351	.279	-.089	-.188	.099	1.000

Fig. 35. Correlation Matrix

For sampling adequacy Kaiser Meyer-Olkin Measure (KMO) test has been used which mainly displays the proportion of variance in the variables. In our analysis our value is barely crossing 0.50 value which is not quite as good.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.594
Bartlett's Test of Sphericity	Approx. Chi-Square	304.832
	df	91
	Sig.	.000

Fig. 36. KMO Bartlett's Test

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.155	29.675	29.675	4.155	29.675	29.675	3.669	26.206	26.206
2	2.681	19.149	48.825	2.681	19.149	48.825	2.801	20.009	46.215
3	1.726	12.326	61.151	1.726	12.326	61.151	1.801	12.864	59.078
4	1.144	8.172	69.323	1.144	8.172	69.323	1.434	10.245	69.323
5	1.000	7.142	76.465						
6	.817	5.833	82.298						
7	.763	5.448	87.746						
8	.510	3.642	91.388						
9	.328	2.341	93.729						
10	.313	2.239	95.968						
11	.214	1.527	97.495						
12	.160	1.141	98.636						
13	.134	.959	99.595						
14	.057	.405	100.000						

Extraction Method: Principal Component Analysis.

Fig. 37. Total Variance

As displayed in Total Variance table only 4 variables highlighted the appropriate significance and stored as new variables. Also from figure 38, it is made clear from scree plot elbow curve starts near 4th variable and then curve is going almost straight with x axis.

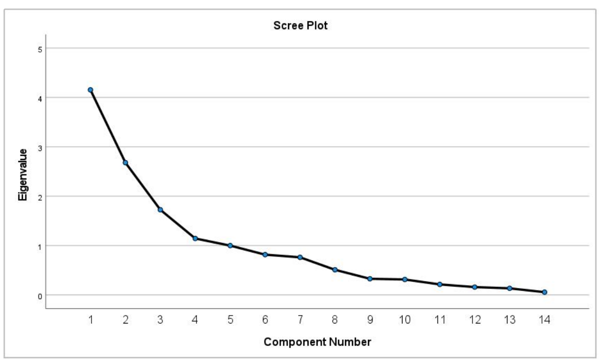

Fig. 38. Scree Plot

Using the 4 newly generated variables , logistic regression has been performed again. The results are displayed below:

In figure 39 the percentage of precision prediction is displayed in Block 0, classification table. The precision value we get is 52.4% .

The Hosmer and Lemeshow Test value is also improved from Model 1 as it went towards 1.

The overall predict percentage that model 2 achieved is 78.6%.

As it can be noticed in figure 42 the significance value of variable factor score 3 is coming as 0.221 which is higher than 0.05 significance level. So in our next model we will be excluding the factor score 3.

Classification Table[a,b]

	Observed		Predicted: smoker 0	Predicted: smoker 1	Percentage Correct
Step 0	smoker	0	0	20	.0
		1	0	22	100.0
	Overall Percentage				52.4

a. Constant is included in the model.

b. The cut value is .500

Fig. 39. Block 0 Classification Table for PCA variables

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	34.961[a]	.424	.566

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

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.612	8	.798

Fig. 40. After fitting model Summary and Hosmer Lemeshow Test

Classification Table[a]

	Observed		Predicted: smoker 0	Predicted: smoker 1	Percentage Correct
Step 1	smoker	0	17	3	85.0
		1	6	16	72.7
	Overall Percentage				78.6

a. The cut value is .500

Fig. 41. Classification table after model fitting at Block 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	REGR factor score 1 for analysis 1	-1.289	.637	4.092	1	.043	.276	.079	.961
	REGR factor score 2 for analysis 1	1.222	.538	5.152	1	.023	3.392	1.181	9.741
	REGR factor score 3 for analysis 1	.479	.391	1.496	1	.221	1.614	.749	3.476
	REGR factor score 4 for analysis 1	1.773	.709	6.244	1	.012	5.888	1.466	23.650
	Constant	.565	.515	1.201	1	.273	1.759		

a. Variable(s) entered on step 1: REGR factor score 1 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 3 for analysis 1, REGR factor score 4 for analysis 1.

Fig. 42. Variable Summary

C. Model-3

While building this model only variables which showed significance level below 0.05 has been taken into consideration and factor score 3 variable has been removed.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	36.489[a]	.403	.537

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.743	8	.459

Fig. 43. After fitting model Summary and Hosmer Lemeshow Test

R2 value of the model fitted is .537 and Hosmer, Lemeshow Test has value as .459 which is lower than the model 2.

Classification Table[a]

			Predicted		
			smoker		Percentage Correct
	Observed		0	1	
Step 1	smoker	0	15	5	75.0
		1	6	16	72.7
	Overall Percentage				73.8

a. The cut value is .500

Fig. 44. Model Summary Hosmer, Lemeshow Test

The overall prediction precision percentage of the model is 73.8% which is quite good as compared with previous model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	REGR factor score 1 for analysis 1	-1.080	.541	3.993	1	.046	.339	.118	.980
	REGR factor score 2 for analysis 1	1.139	.502	5.152	1	.023	3.125	1.168	8.358
	REGR factor score 4 for analysis 1	1.567	.597	6.887	1	.009	4.792	1.487	15.445
	Constant	.417	.459	.827	1	.363	1.518		

a. Variable(s) entered on step 1: REGR factor score 1 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 4 for analysis 1.

Fig. 45. Classification Table after Model Fitting

In our third model the significance of all the 3 variables is 0.046, 0.023 and 0.009 which is quiet lesser than 0.05 significance value.

The Q-Q plot has been plotted for the predicted data. Most of the data seems to be close to the mean line.

V. Summary

In this terminal assessment we have gathered in-depth knowledge about Time Series analysis and Logistic regression. In time series analysis we have performed assessment for

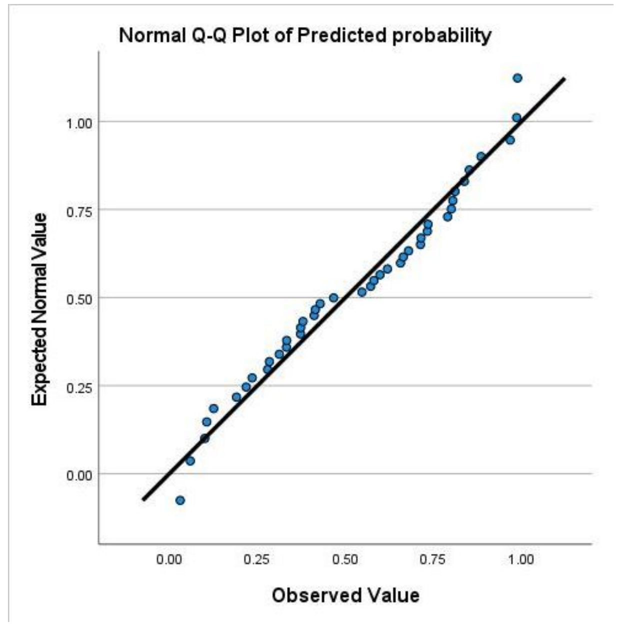

Fig. 46. Q-Q plot for the Predicted data

the components of time series, applied basic, exponential and complex models for each of the time series. By plotting different plots we analysed the trend, seasonality of the data. Understanding ACF and PACF graphs as well as differencing of the data were also knowledge domain in this assessment. We gained knowledge about mean, naive, seasonal naive, ETS, Holt winters, ARIMA, Seasonal ARIMA models. To apply a particular model basic requirement for the model were satisfied. Whereas in Logistic Regression we have build three different models by analysing various factors and taking all the necessacry steps into account. We also performed Principal Component Analysis for the data as it seemed to be necessary step in getting significant results.

References

[1] Forecasting: Principles and Practice [Online].Available: https://otexts.com/fpp2/

[2] An Approach to Time Series Analysis [Online].Available: https://doi.org/10.1214/aoms/1177704840/

[3] An Introduction to Logistic Regression Analysis and Reporting [Online].Available: https://www.tandfonline.com/doi/abs/10.1080/00220670209598786/