TABA Time Series, Logistic Regression and Principal Component Analysis using R

Sachin Muttappanavar Student Id: x20144253
MSc in Data Analytics
National College of Ireland Dublin, IRELAND
URL: www.ncirl.ie

Abstract: **In this project paper we have implemented three statistical techniques – Time series analysis, Logistic Regression and Principal component analysis.**
Time series analysis technique is applied to forecast overseas trips and to forecast new house registration that are expected occur in the next two periods. We have implemented different time series models like seasonal naïve, exponential smoothing, ARIMA models. We evaluated each model to select best model based on metrics like RMSE, MAPE, AIC. We were able to forecast overseas trips and new house registration more accurately using time series technique. Logistic regression method is applied on child births dataset to classify whether newly born child is having low eight or not. Principal component analysis technique is applied to transform higher dimensional data into lower dimension. We also tuned threshold value for logistic regression to get optimal value for sensitivity and specificity. Confusion matrix, kappa value, accuracy, sensitivity, and specificity value are interpreted for each model to find best model. Using Logistic Regression technique, we were able to classify childbirth weight more accurately.

I. OBSERVATIONS

Time Series: Time series is sequence of data points with distinct-time period.

Seasonal Naive: This method is used for highly seasonal time series data. In this method forecast values will be equal to the last observed value from the same season of the year.

Simple exponential smoothing: It is forecasting method that can fit time series data consisting constant level. Alpha is the weight given to previous observation. Alpha value ranges between 0 to 1. It decreases exponentially. High weight is given to most recent observation and weight decreases exponentially for observation coming from further in the past. In case of SES, forecast will be equal to last level component.
Level Equation:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

Holt: This method of forecasting is suitable for time series data which consists of level and trend component. Here, Beta is the trend smoothing parameter. It ranges from 0 to 1. Most recent observation will be given high weight and decreases exponentially further down.

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + h b_t \\ \text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*) b_{t-1}, \end{aligned}$$

Holt-Winters: This method of forecasting technique consists of three components like level, trend and seasonal. Here, Gamma value is the seasonal smoothing parameter. It ranges from 0 to 1. Recent observation is given more weight and weight decays exponentially further down.

$$\begin{aligned} \text{level} \quad L_t &= \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}); \\ \text{trend} \quad b_t &= \beta(L_t - L_{t-1}) + (1-\beta) b_{t-1}, \\ \text{seasonal} \quad S_t &= \gamma(y_t - L_t) + (1-\gamma) S_{t-s} \\ \text{forecast} \quad F_{t+k} &= L_t + k b_t + S_{t+k-s}, \end{aligned}$$

ARIMA : ARIMA is abbreviate for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

II. MODELS BUILDING PROCESS AND DESCRIPTION:

Overseas Trips:
Dataset Description: Data is provided in the csv file format. This data gives information about non-residents overseas trips to Ireland from quarter 1 of 2012 to quarter 4 of 2019. We have used R language to work on this time series data. Data looks like below:

```
   ï..Quarter Trips.Thousands.
1      2012Q1           1165.1
2      2012Q2           1817.3
3      2012Q3           2096.7
4      2012Q4           1438.0
5      2013Q1           1251.7
6      2013Q2           1893.0
7      2013Q3           2261.0
8      2013Q4           1580.1
9      2014Q1           1342.5
10     2014Q2           2126.6
```

We have created time series data in R as follow:

```
dt<-ts(over_sea_trips_ts$Trips.Thousands.,start = c(2012,1), frequency = 4)
```

Time series data is decomposed to separate seasonal, trend, irregular component from time series data. Plot of decomposed data looks like below:

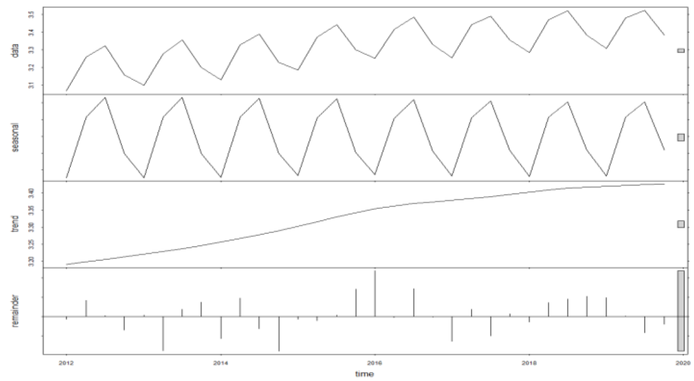

Figure 1. decomposed timeseries

In the above plot we can see clearly that seasonal and trend components are significant as error bar is significant which is on right side of graph, whereas remainder plot is not significant as lines are within error bar. This depicts time series data consists of seasonal as well as trend in it.

Models: We have explored different time series technique from simple naïve method to SARIMA model.

Model 1: First we have built seasonal naïve model as our data consists of seasonal component. Below plot shows forecast from seasonal naïve method appended to original time series data. Blue line is forecast value on an average. Greyed out area around blue line shows forecast value with 80 percent and 95 percent confidence interval.

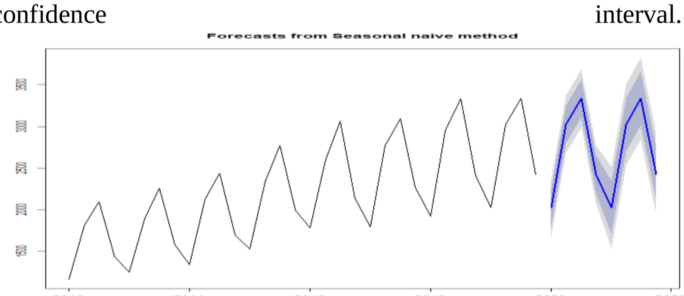

Figure 2. Forecast from seasonal naive.

Seasonal naïve model properties are shown in below Figure 3. Quadric mean of model predicted values and actual values is 178.65 with mean absolute percentage error of 6.975.

```
Forecast method: Seasonal naive method

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

Residual sd: 176.6505

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

Figure 3. Summary of seasonal naive model

```
        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
```

Model 2: Holt winters model consists of three smoothing parameters (one for level, one for trend and one for seasonal component) and forecast equation. Below is the forecast produced by holt winters model.

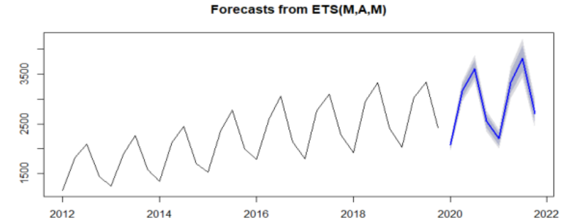

Figure 4. Forecast from ETS(M,A,M)

Model summary: ETS function of forecast package is used to fit holt winter model. We have used Z value to automatically select additive or multiplicative type for model. We tried to pick best values for smoothing parameter by building models with different values for smoothing parameter. Smoothing parameters with alpha = 0.7, beta = 0.0119 and gamma = 10^ (-4) gave good results in terms of RMSE (53.97879) and MAPE (1.998315).

```
ETS(M,A,M)

Call:
 ets(y = dt, model = "ZZZ", alpha = 0.7, beta = 0.0011, gamma = 1e-04)

  Smoothing parameters:
    alpha = 0.7
    beta  = 0.0011
    gamma = 1e-04

  Initial states:
    l = 1529.1177
    b = 39.5596
    s = 0.8793 1.2603 1.1156 0.7448

  sigma:  0.0267

     AIC     AICc      BIC
372.2793 375.6393 381.0737

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set -2.841834 53.97879 44.51095 -0.09613793 1.998315 0.2904873 -0.08387628
```

Forecast result using model is shown in below Figure5.

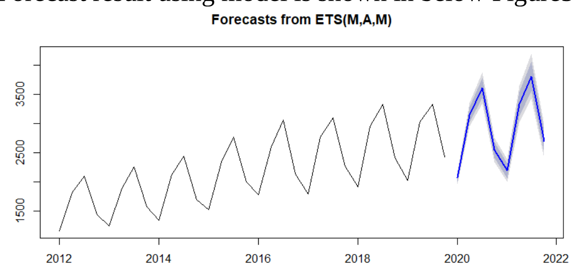

Figure 5. Forecast by Holt winters

```
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2075.130 2004.074 2146.186 1966.459 2183.801
2020 Q2       3152.175 3020.943 3283.406 2951.473 3352.876
2020 Q3       3611.019 3438.567 3783.471 3347.277 3874.762
2020 Q4       2554.113 2418.583 2689.644 2346.837 2761.389
2021 Q1       2192.752 2066.045 2319.460 1998.970 2386.535
2021 Q2       3328.350 3121.803 3534.897 3012.464 3644.236
2021 Q3       3810.058 3558.698 4061.418 3425.636 4194.480
2021 Q4       2692.982 2505.559 2880.405 2406.344 2979.620
```

Model 3: We next built bit more advanced time series model ARIMA. As our time series data has seasonal component, we have fitted SARIMA model. To implement this model we leveraged auto.arima function of forecast library. Summary of auto.arima model is as follow:

```
Series: dt
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

Training set error measures:
                   ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set 1.570482 67.54754 55.32141 -0.09590574 2.451533 0.3610385 -0.0396672
```

Auto.arima() function selected p=1, d=0, q=0 for non-seasonal part and P=0,D=1,Q=0 at lag 4 for seasonal part. To check residuals from model are all zero we conducted Ljung-Box test. Null hypothesis for this test is all autocorrelations of residuals are zero.

Results displayed in below Figure 6.

```
Ljung-Box test

data:  Residuals from ARIMA(1,0,0)(0,1,0)[4] with drift
Q* = 4.277, df = 4, p-value = 0.3698

Model df: 2.   Total lags used: 6
```

Figure 6. Ljung-Box test

We are failed to reject null hypothesis as p-value is >0.05. Therefore, autocorrelations of residuals are all zero. We can also refer to acf plot shown in Figure 7.

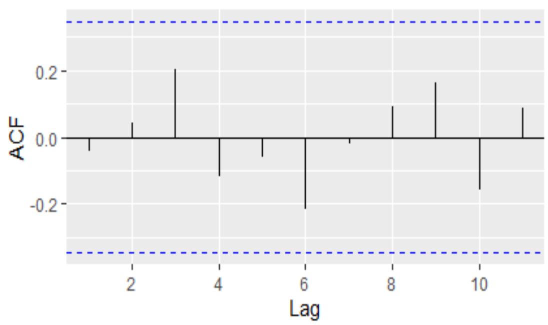

Figure 7. Autocorrelation plot

Figure 8 shows residuals are normally distributed with mean zero.

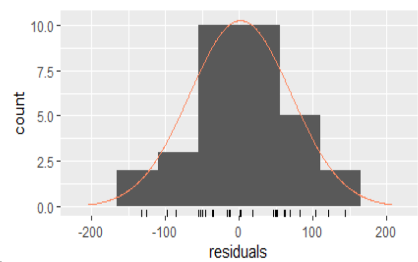

Figure 8. Histogram of residuals

```
        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
2020 Q4       2553.069 2435.612 2670.525 2373.435 2732.703
2021 Q1       2228.305 2069.305 2387.305 1985.135 2471.474
2021 Q2       3259.194 3088.339 3430.048 2997.894 3520.493
2021 Q3       3592.676 3417.969 3767.383 3325.485 3859.867
2021 Q4       2695.061 2519.062 2871.060 2425.893 2964.229
```

Final Model for Overseas Trips: In total we built three models of different family. First, we built simple time series model of seasonal naive. This model is making comparatively large error in prediction than other two models. RMSE and MAPE value is bit higher than other models.

SARIMA model can predict forecast more accurately than seasonal naïve model. But when compared to holt winters model values are high with respect to RMSE and MAPE.
Holt-Winters giving good results with reasonable RMSE value of 53.97879 and MAPE value of 1. 99831.Thus, Holt-Winters model is selected as best model for forecasting overseas trips in next two years.

New House Registration:
Dataset Description: This dataset comprises information about annual series of new house registration from 1978 to 2019. R language is used to analyze this time series dataset.
As dataset consists of only yearly data and no monthly information, dataset does not have seasonal component. So, we have built 3 different non seasonal models and compared to find best model among them. Figure 9 shows plot of time series data with years on x -axis and new house registration count on y-axis.

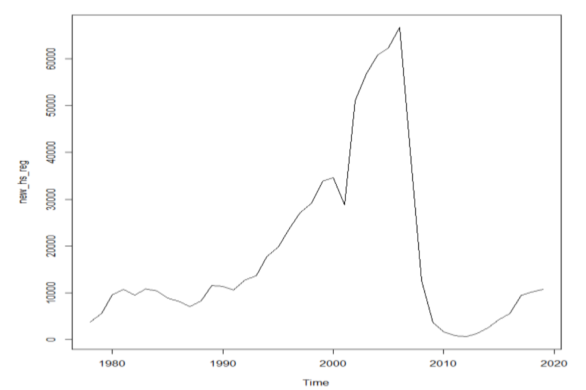

Figure 9. New house registration time series data

Model 1: Holt model is suitable for time series data with linear trend and no seasonal component in it. As our data consists of trend component alone, we implemented Holt model with help of ets() function of forecast linrary. We have built two holt models, with and without damp.

Summary of Holt with damp:

```
ETS(M,Ad,N)

Call:
 ets(y = new_hs_reg_holt, model = "ZZN", 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
```

Summary of Holt model without damp:

```
ETS(M,A,N)

Call:
 ets(y = new_hs_reg_holt, model = "ZZN", 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
```

Both models are using alpha = 0.9999, beta = 0.001 as smoothing parameter for exponential decay of the level and decay of trend, respectively. AIC value of both models are compared to select best model. Model without damp is having lower AIC value with reasonable RMSE and MAPE value. So, choosing model without damp from these two models. Below is the graph of models forecast.

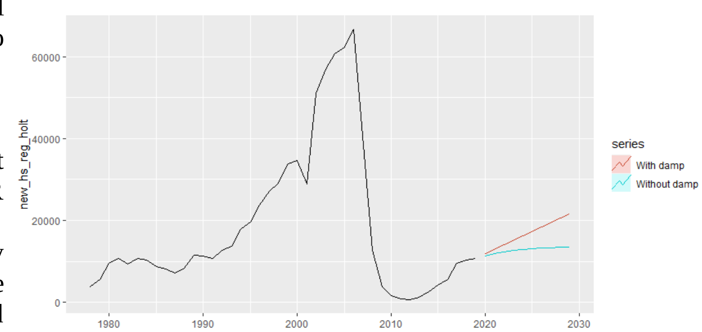

Figure 10. Forecast by holt models

Arima models: We have built two ARIMA models. For first model we have picked (p,d,q) values manually and in later one we have used auto.arima function to select these values automatically by machine.

Model 2: Using ndiffs() function we try to find number of differences required to make time series stationary. Function returned '0', that means time series is already stationary. We also conducted kpss test to check whether the time series data is stationary or not. As p-value is lesser than 0.05 null hypothesis is rejected. That means trend is stationary.

```
        KPSS Test for Trend Stationarity

data:  new_hs_reg
KPSS Trend = 0.17419, Truncation lag parameter = 3, p-value = 0.02651
```

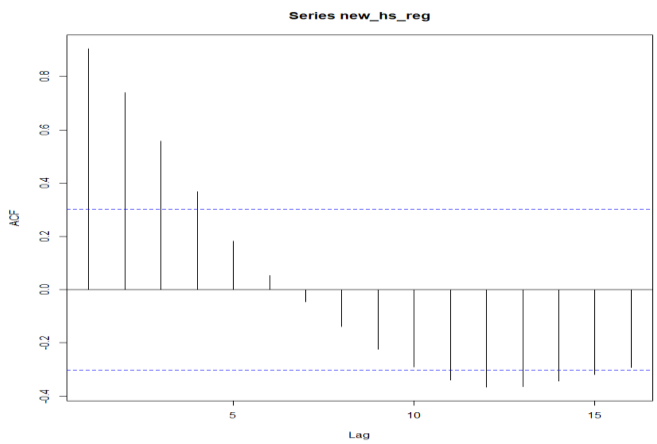

Figure 11. Auto correlation plot

From above Autocorrelation plot we can see that there are 4 significant lines. So, we are assigning q = 4.

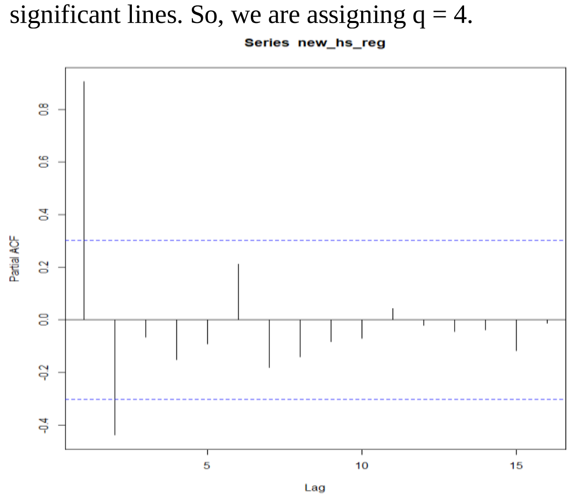

Figure 12. partial auto correlation

There are two significant lines at lag 1 and 2 in the above pacf plot. So, assigning p = 2. Finally we built ARIMA model with (p, d,q) = (2,0,4) and its summary is shown below:

```
Series: new_hs_reg
ARIMA(2,0,4) with non-zero mean

Coefficients:
         ar1      ar2     ma1     ma2     ma3     ma4       mean
      0.6576  -0.0850  0.6906  0.5916  0.4908  0.4789  16724.008
s.e.  0.2932   0.2706  0.2549  0.2568  0.2561  0.1692   6507.684

sigma^2 estimated as 41453024:  log likelihood=-425.68
AIC=867.37   AICc=871.73   BIC=881.27

Training set error measures:
                   ME    RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 182.9685 5877.43 3473.094 -28.92229 63.62506 0.8754425 0.02195582
```

Ljung-Box test is conducted to check the residuals from ARIMA model are with zero mean. Probability value is greater than 0.05, that means residuals are having zero autocorrelation at every lag (Figure 13).

```
        Ljung-Box test

data:  Residuals from ARIMA(2,0,4) with non-zero mean
Q* = 1.1985, df = 3, p-value = 0.7534

Model df: 7.   Total lags used: 10
```

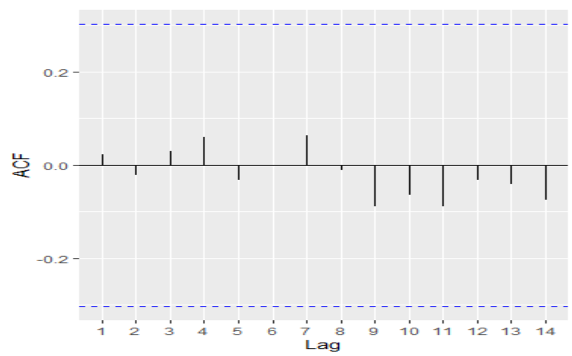

Figure 13. Autocorrelation of the residuals

We can also see in the Figure that residuals are normally distributed. Hence ARIMA model is appropriate.

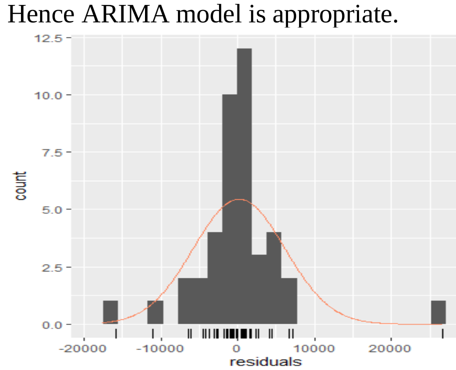

Figure 14. Histogram of residuals

Here we have visualized forecast prediction of the ARIMA model in the Figure 15.

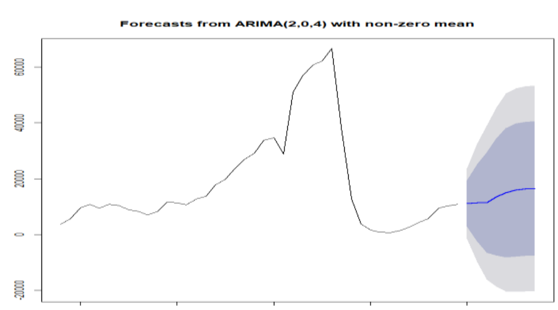

Figure 15. Forecast from ARIMA(2,0,4)

Model 3: This ARIMA model is built using auto.arima() function of the forecast package. Model built with this function has p, d, q value as 2 ,0 ,0. Order of AR is 2 with co efficient 1.3346, -0.4665. Order of MA is zero and number of differencing required is zero.

Model summary is as below:

```
Series: new_hs_reg
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
```

Forecast prediction by model3 is shown in Figure 16.

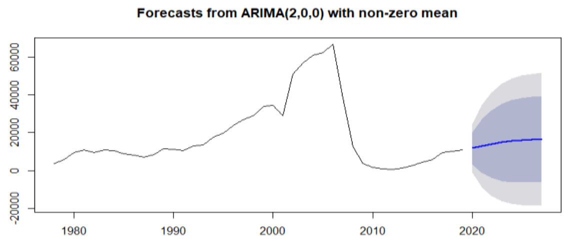

Figure 16. Forecast from ARIMA(2,0,0)

To check whether built model is appropriate or not, we examined residuals autocorrelations and distribution.
From Figure 17, autocorrelations of residuals are all zero as line at each lag are within blue dash line

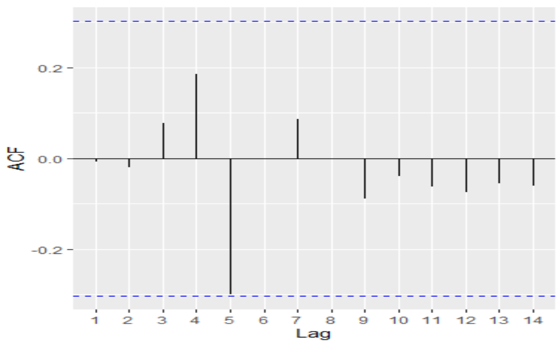

Figure 17. Autocorrelation of residuals ARIMA(2,0,4)

Distribution of Residuals (Figure 18) are checked and are normally distributed around zero. Therfore, model is appropriate.

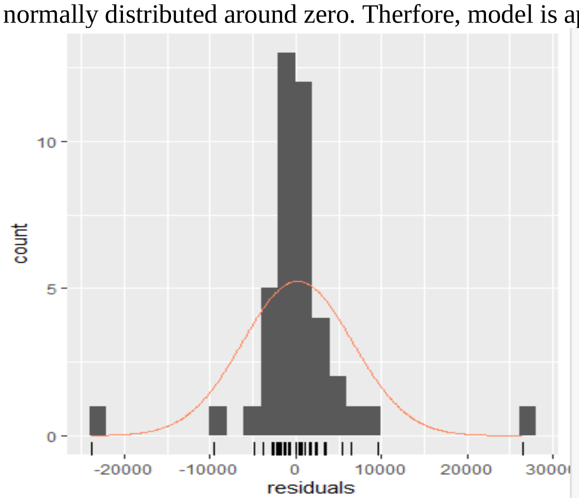

Figure 18. histogram of residuals ARIMA(2,0,4)

We have considered AIC value for comparison as d value is same and models are from same family. Among two ARIMA models, model 3 is having least AIC value. So, selecting ARIMA (2,0,0) from ARIMA models. RMSE and MAPE value for ARIMA (2,0,0) are 6342.208, 35.95662 Forecast values are as below.

```
     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
2023       14846.94  -5450.112 35144.00 -16194.723 45888.61
2024       15501.24  -6161.059 37163.55 -17628.390 48630.88
2025       15976.65  -6409.636 38362.94 -18260.221 50213.52
2026       16305.88  -6435.970 39047.74 -18474.780 51086.55
2027       16523.49  -6379.445 39426.43 -18503.527 51550.51
```

Final Model for New House registration: Seasonal Exponential smoothing model is a simple model and forecast values are same as recent observation. It is capturing only level parameters. So, we are rejecting this model.

We are considering RMSE score and MAPE value for comparing models. Holt model is making high root mean square error and mean absolute percentage error. Whereas, for ARIMA model, quadratic mean of difference between actual and predicted value is reasonable compared to other models. We have performed all diagnostic test for ARIMA (2,0,0) and model is appropriate for forecasting.

So, ARIMA (2,0,0) model is adequate to forecast new house registration in Ireland for next three periods.

Child Births:
Dataset Description: Dataset contains information about child births in US city. In this work, we have used Logistic Regression statistical model to find whether newly born baby has low birth weight or not. Target variable in the dataset is 'lowbwt' (0 = No, Positive and 1 = yes, Negative). Datasets consists of 42 rows and 16 columns.

Information about dataset:

```
'data.frame':   42 obs. of  16 variables:
 $ ï..ID       : int  1360 1016 462 1187 553 1636 820 1191 1081 822 ...
 $ Length      : int  56 53 58 53 54 51 52 53 54 50 ...
 $ Birthweight: num  4.55 4.32 4.1 4.07 3.94 3.93 3.77 3.65 3.63 3.42 ...
 $ Headcirc    : int  34 36 39 38 37 38 34 33 38 35 ...
 $ Gestation   : int  44 40 41 44 42 38 40 42 38 38 ...
 $ smoker      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mage        : int  20 19 35 20 24 29 24 21 18 20 ...
 $ mnocig      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mheight     : int  162 171 172 174 175 165 157 165 172 157 ...
 $ mppwt       : int  57 62 58 68 66 61 50 61 50 48 ...
 $ fage        : int  23 19 31 26 30 31 31 21 20 22 ...
 $ fedyrs      : int  10 12 16 14 12 16 16 10 12 14 ...
 $ fnocig      : int  35 0 25 25 0 0 0 25 7 0 ...
 $ fheight     : int  179 183 185 189 184 180 173 185 172 179 ...
 $ lowbwt      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mage35      : int  0 0 1 0 0 0 0 0 0 0 ...
```

Figure 19. Dataset information

As we can see in the above Figure 19 smoker, lowbwt, mage35 are holding categorical values but type of column is int. We converted these columns into factor.
In the below Figure 20 we have visualized correlation between numerical values in the dataset. We did not find any input variables which are highly correlated with other input variable.

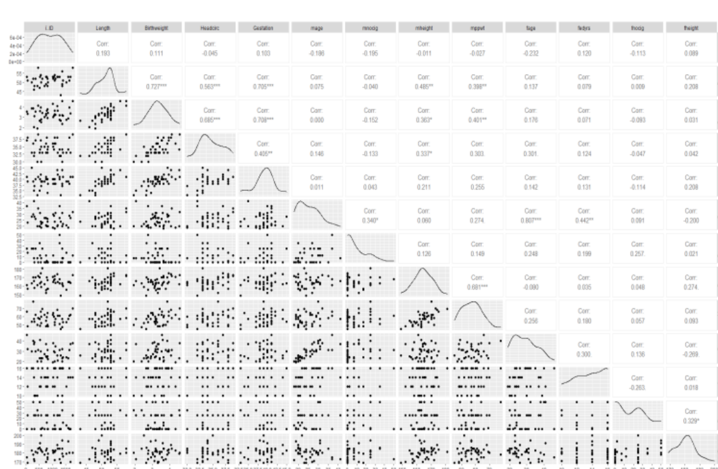

Figure 20. Correlation plot

Then we explored each numerical variable with respect to dichotomous target variable. In the below Figure 21 we have visualized numerical values variation with respect to target variable.

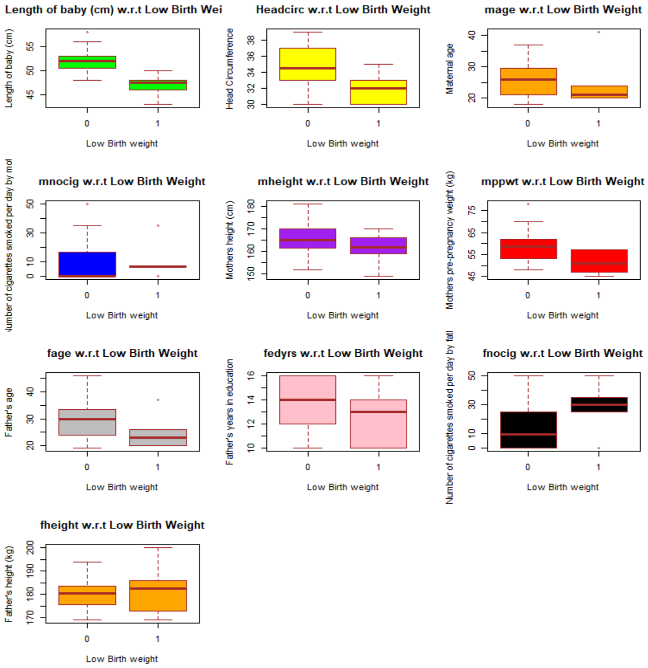

Figure 21. Box plot of numerical values w.r.t low birth weight

We also conducted statistical hypothesis test to check likelihood probability of given sample data belongs to population data. We conducted ANOVA test on numerical input variables and target variables which is categorical. Levenes Test is carried out before ANOVA test to check population variance are equal. p-value is greater than 0.05 thus null hypothesis (groups have equal population variances) is failed to reject. For categorical input variables, chi square statistical test is conducted. Results of test is shown in below Table 1.

Variable	**Test**	**p-value**
Length	ANOVA	1.8e-05
Headcirc	ANOVA	0.00301
mage	ANOVA	0.631
mnocig	ANOVA	0.824
mheight	ANOVA	0.208
mppwt	ANOVA	0.0215
fage	ANOVA	0.118
fedyrs	ANOVA	0.225
fnocig	ANOVA	0.0886 .
fheight	ANOVA	0.534
smoker	Chi-Square	0.9976
mage35	Chi-Square	0.6047

Table 1. Results of hypothesis test

Variables which are having significant value less than 0.05 are used for building model.

Models & Summary:

Model 1: We have selected Length, Headcirc, mppwt as input variables as they have p-value less than 0.05. Logistic regression formula for Model 1 is like below:

```
glm(formula = lowbwt ~ Length + Headcirc + mppwt, family = binomial,
    data = df)
```

Coefficients and significance of input variables are as below:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  90.8521    39.5203   2.299   0.0215 *
Length       -1.3271     0.6964  -1.906   0.0567 .
Headcirc     -0.3887     0.5073  -0.766   0.4435
mppwt        -0.2630     0.1974  -1.332   0.1828
```

As we can see none of the input variables are significant. So rejecting this model.

Model 2: We have built different models with different combination of input variables, among all model with Length and headcirc interaction as input variable produced good results, even they are significant in predicting target variable. Model summary is like below:

```
Call:
glm(formula = lowbwt ~ Length:Headcirc, family = binomial, data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.44336  -0.24486  -0.12878  -0.02884   2.66015

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     25.297244   9.450418   2.677  0.00743 **
Length:Headcirc -0.016461   0.005948  -2.767  0.00565 **
```

Confusion matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 35  2
         1  1  4

               Accuracy : 0.9286
                 95% CI : (0.8052, 0.985)
    No Information Rate : 0.8571
    P-Value [Acc > NIR] : 0.1312

                  Kappa : 0.6866

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.9722
            Specificity : 0.6667
         Pos Pred Value : 0.9459
         Neg Pred Value : 0.8000
             Prevalence : 0.8571
         Detection Rate : 0.8333
   Detection Prevalence : 0.8810
      Balanced Accuracy : 0.8194

       'Positive' Class : 0
```

Accuracy of our model with default (0.5) threshold is 0.9286. When we look at specificity, it is low, that means model is not good in predicting negative cases. It is bad to classify normal baby as low weight. So, we tried to increase specificity by adjusting threshold value. We tried several threshold values and found that threshold value of 0.4 is giving better results.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 35  1
         1  1  5

               Accuracy : 0.9524
                 95% CI : (0.8384, 0.9942)
    No Information Rate : 0.8571
    P-Value [Acc > NIR] : 0.04923

                  Kappa : 0.8056

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9722
            Specificity : 0.8333
         Pos Pred Value : 0.9722
         Neg Pred Value : 0.8333
             Prevalence : 0.8571
         Detection Rate : 0.8333
   Detection Prevalence : 0.8571
      Balanced Accuracy : 0.9028

       'Positive' Class : 0
```

Now model looks perfect with better negative prediction rate. Kappa value is good that indicates performance of model is better. Accuracy of model is 95.24 %. ROC curve of model is visualized in Figure 22. Area under curve is 93.3 % that indicates model performance is better.

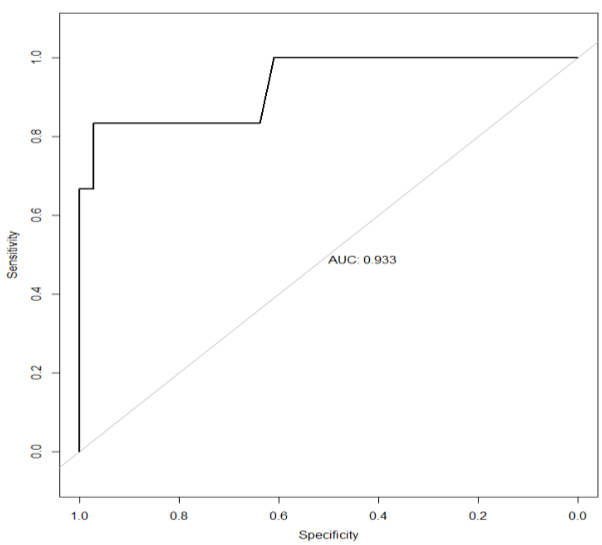

Figure 22. ROC curve

Model 3: We applied principal component analysis technique to transform data from higher to lower dimension thereby addressing curse of dimensionality. We have visualized number of components Vs eigen values of components in below scree plot. As per Catell's scree test we are retaining the components above the level off point or elbow in the plot.

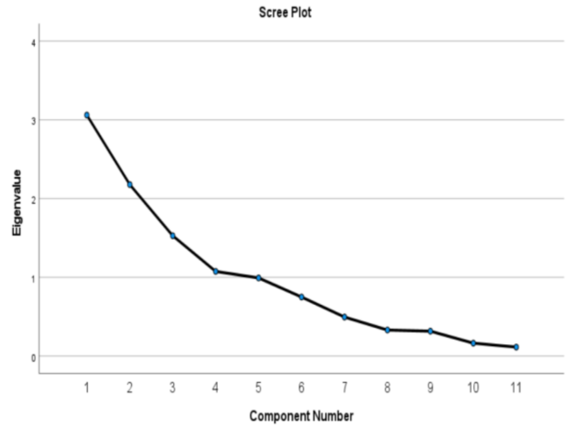

Figure 23. Scree plot

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SS loadings	3.06	2.27	1.58	1.14	1.05	0.81	0.73	0.48	0.33	0.28
Proportion Var	0.26	0.19	0.13	0.10	0.09	0.07	0.06	0.04	0.03	0.02
Cumulative Var	0.26	0.44	0.58	0.67	0.76	0.83	0.89	0.93	0.95	0.98
Proportion Explained	0.26	0.19	0.13	0.10	0.09	0.07	0.06	0.04	0.03	0.02
Cumulative Proportion	0.26	0.45	0.59	0.69	0.78	0.85	0.91	0.95	0.98	1.00

There is a distinct division between first three eigen values and other. So, retaining first three components for model building. First three components can explain 58 % variance in the data. We built various models with varying threshold, among all model with 0.5 threshold gave best results.

```
Prediction  0  1
         0 33  1
         1  3  5

               Accuracy : 0.9048
                 95% CI : (0.7738, 0.9734)
    No Information Rate : 0.8571
    P-Value [Acc > NIR] : 0.2644

                  Kappa : 0.6585

 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.9167
            Specificity : 0.8333
         Pos Pred Value : 0.9706
         Neg Pred Value : 0.6250
             Prevalence : 0.8571
         Detection Rate : 0.7857
   Detection Prevalence : 0.8095
      Balanced Accuracy : 0.8750

       'Positive' Class : 0
```

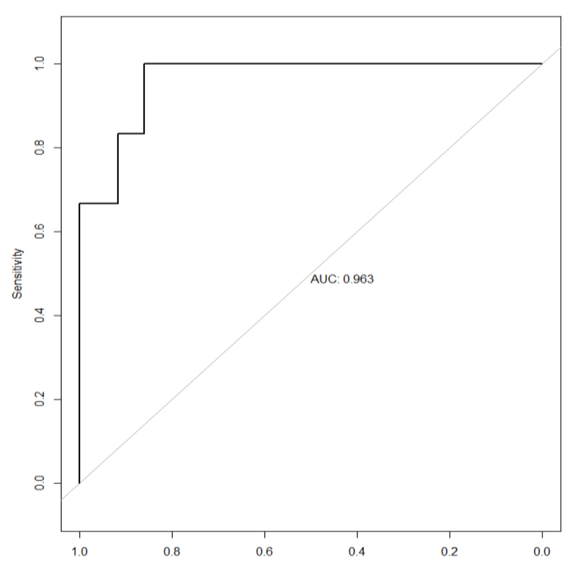

Figure 24. ROC curve

This model is having low kappa value that indicates models is not classifying minority classes correctly.

Final Model for Child Births: Among all models, model 2 is simple and showing good results in terms Accuracy, AUC, Sensitivity, Specificity with least number of input variables. So, Model 2 will be best model to predict low birth weight of child.

III.CONCLUSION

We developed exponential, ARIMA models to forecast time series data. Using Holt winters method, we were able to forecast Overseas trips more accurately. For forecasting new house registration dataset, ARIMA (2,0,0) model is suitable as it is making less error in forecasting and values are more accurate than other models. Logistic regression with length and head circumference input variable was able to classify the child with low birth weight with high accuracy. We also explored Principal Component Analysis technique and applied to childbirth dataset to transform data into lower dimension.

REFERENCES

[1] M. Ivanović and V. Kurbalija, "Time series analysis and possible applications," 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016, pp. 473-479, doi: 10.1109/MIPRO.2016.7522190.
[2] J. Contreras, R. Espinola, F. J. Nogales and A. J. Conejo, "ARIMA models to predict next-day electricity prices," in IEEE Transactions on Power Systems, vol. 18, no. 3, pp. 1014-1020, Aug. 2003, doi: 10.1109/TPWRS.2002.804943.
[3] Practical Statistic for Data Scientist by orielly book.
[4] Early Prediction of LBW Cases via Minimum Error Rate Classifier: A Statistical Machine Learning Approach. - Gloria Miró Amarante, Victoria E. Rey Caballero.
[5] X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.