Statistical Analysis of Time Series, Logistic Regression and Principal Component Analysis

Akash Manjunatha
School of Computing
National College of Ireland
Dublin, Ireland
x21141797@student.ncirl.ie

***Abstract*— In this paper, an attempt is made to perform time series forecasting on a dataset of private car registrations in Ireland from January 1995 to January 2022, and binary logistic regression and principal component analysis are used to predict whether customers have default history or not, using the historical data R program and SPSS.**

I. OBJECTIVES OF TIME SERIES

The purpose of this project is to determine the best time series for the dataset and then apply the chosen time series model. To do so, many comparisons such as Root mean square error (RMSE) and Akaike information criterion (AIC) is considered. Finally, the most appropriate model is picked and forecasted for the next six periods.

A. Description of the Dataset

The data set is a monthly time series of private car registrations in Ireland from January 1995 to January 2022, sourced from the Central Statistics Office of Ireland. The file is in CSV format, and the data description is as follows: Year, Month, and the number of cars registered.

B. Steps followed to build the model

Step 1: Identifying the pattern – The primary aim of time series(TS) data is to find the pattern of the given dataset. When the available data set is plotted against the year in Fig. 1, in the preliminary analysis it looks like data is not stationary; some fluctuations can be seen, with a dip in 2010 and a linear increase in registration over the next five years. Because this graph is made up of 26 years of data, it is difficult to evaluate or draw conclusions from it in terms of seasonality over time. Seasonal plot Fig. 2 and seasonal subseries plot Fig. 3 are plotted to get a clear picture and to identify the presence of seasonality and trend.

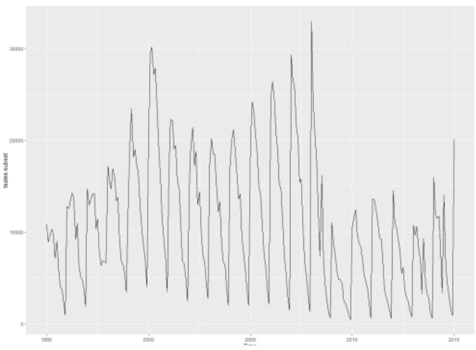

Fig. 1. Time Series Pattern.

The seasonal plots depict that, on average, the maximum number of automobiles are registered in January, while the lowest number of cars are registered in the last month of the year. However, until mid-year (June), the registration of cars decreases linearly, with a sudden spike in July and a sudden decrease in the following month, returning to where it was before last month (possibly due to the transition from summer to winter), and gradually registrants decrease linearly, reaching the lowest in December. By this, we can say that our data set has seasonality and trend.

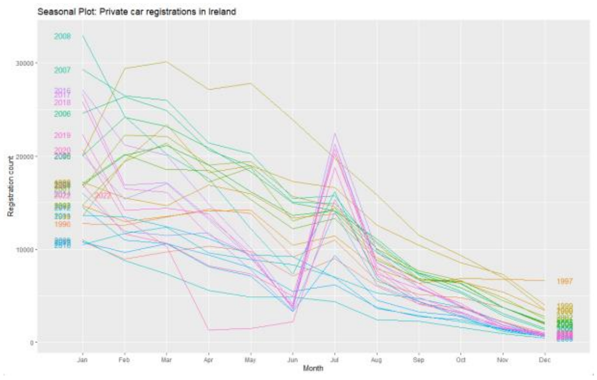

Fig. 2. Monthly Car Registration Plot.

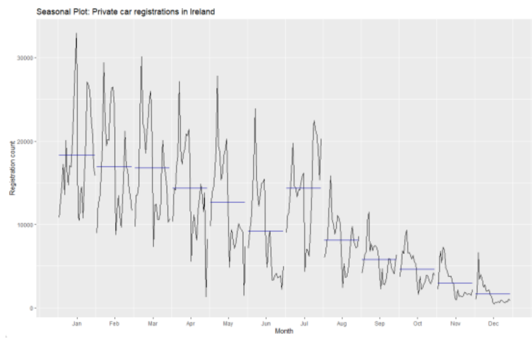

Fig. 3. Seasonal Subseries Plot.

Step 2: Seasonal Decomposition The presented data has a seasonality, which may be broken down into three components: seasonal, trend and irregular.

-Trend: captures level over time.

-Seasonal component: catches over the time of years.

-Irregular component: captures patterns not recorded in the first and second components by examining Figs. 1 and 2 above. We can say that our data is non-linear because we see more car registration count at the beginning of the year and a decrease in the number of registrations each month at the end of the year. We can conclude that our data is multiplicative decomposition [1]

A multiplicative model is calculated by the given equation Y t = Trend t * Seasonal t * Irregular t, Fig. 4 decomposed plot depicts the multiplicative decomposition

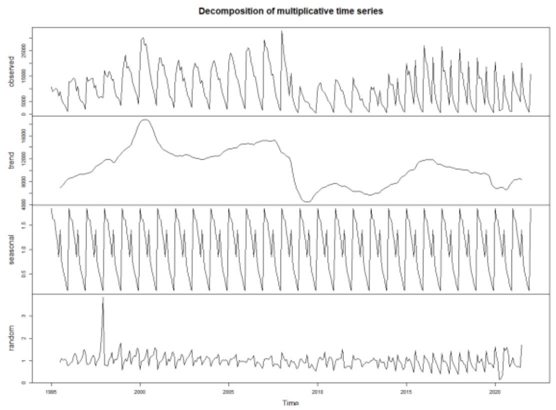

Fig. 4. Multiplicative Time Series Plot (Decomposed).

C. Model build and Forecast:

1) Model 1: There are many simple methods of forecasting time series data available, that includes Mean, Naïve, Seasonal Naïve, and Drift, among which we chose Seasonal Naïve. As shown in Fig. 1 and 2, our model has a seasonal trend that decreases linearly in the beginning, increases in the middle, and then decreases again. We are also forecasting for the next six months, and the best way to do so is to use seasonal Naïve, which uses the last observation of the previous month's forecasted period. To double-check the model, it was run through other methods as well, and the Seasonal RMSE value was determined to be the lowest of 3475.13 among others, confirming our hypothesis. Below are the results and a projected plot of the model.

```
> summary(fcast.seasonalnaive)

Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = ts2, h = 6)

Residual sd: 3475.1351

Error measures:
                   ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 57.47284 3475.135 2184.406 -9.056944 29.09186    1 0.7196226

Forecasts:
         Point Forecast      Lo 80     Hi 80      Lo 95    Hi 95
Feb 2022          11672  7218.4351 16125.565  4860.8603 18483.14
Mar 2022          10672  6218.4351 15125.565  3860.8603 17483.14
Apr 2022           8214  3760.4351 12667.565  1402.8603 15025.14
May 2022           7337  2883.4351 11790.565   525.8603 14148.14
Jun 2022           4980   526.4351  9433.565 -1831.1397 11791.14
Jul 2022          20232 15778.4351 24685.565 13420.8603 27043.14
```

Fig. 5. Seasonal Naïve Results.

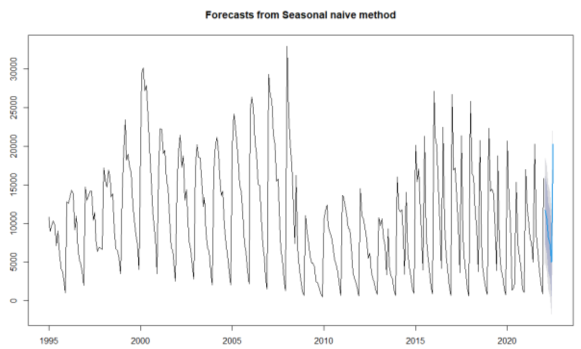

Fig. 6. Seasonal Naïve 6 period forecast Plot.

2) Model 2: Exponential smoothening model (ESM) is a well-known and straightforward approach to forecasting that produces good results in the short run. Because we have a level, trend, and seasonal(alpha, beta, and gamma) component in our model, we employed a triple exponential model, often known as Holt-winters exponential smoothing. This method has two forms depending on the nature of the plot as ours is multiplicative. The multiplicative method is carried out and also to cross-check additive method is run the results of the model can be found in Fig. 7 and 8

```
Holt-Winters' additive method

Call:
 hw(y = ts2, seasonal = "additive")

  Smoothing parameters:
    alpha = 0.4944
    beta  = 0.0028
    gamma = 0.465

  Initial states:
    l = 5420.9066
    b = 242.1843
    s = -8900.013 -7547.437 -5896.908 -4878.688 -2510.886 3842.064
           -1303.185 2139.984 3904.72 6422.741 6636.091 8091.517

  sigma:  2671.182

     AIC     AICc      BIC
7026.015 7028.009 7090.340
> accuracy(hwFit1)
                    ME   RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -226.5401 2604.6 1820.253 -2.517114 32.13705 0.8332943 0.2135645
```

Fig. 7. Additive method result (Holt-Winters).

```
Holt-Winters' multiplicative method

Call:
 hw(y = ts2, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.2556
    beta  = 0.003
    gamma = 0.4619

  Initial states:
    l = 6121.9795
    b = 245.974
    s = 0.8178 0.5261 0.2023 0.6816 0.7986 1.1622
           1.0662 1.3483 1.4575 1.3594 1.325 1.2551

  sigma:  0.2437

     AIC     AICc      BIC
6815.197 6817.191 6879.522
> accuracy(hwFit2)
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -141.3668 2291.275 1533.079 -10.50014 22.14336 0.7018289 0.3422716
```

Fig. 8. Multiplicative method result (Holt-Winters).

The additive and multiplicative RMSE values are 2604.6 and 2291.27, respectively, with AIC values of 7026.01 and 6815.19. We can conclude from the least RMSE and AIC that the multiplicative approach worked better. The results of both are plotted in Fig. 9, and the best model's forecast for the next six months is shown in Fig. 10.

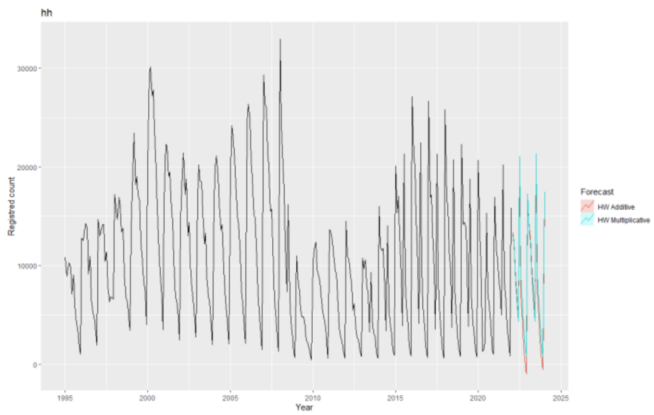

Fig. 9. Combined additive and Multiplicative Forecast Plot.

```
> forecast(hwFit2,h=6)
         Point Forecast       Lo 80      Hi 80     Lo 95      Hi 95
Feb 2022      13134.256   9032.917 17235.596 6861.798 19406.714
Mar 2022      12274.437   8308.271 16240.602 6208.710 18340.163
Apr 2022       8287.884   5520.542 11055.226 4055.600 12520.169
May 2022       6710.578   4398.051  9023.104 3173.874 10247.281
Jun 2022       4325.104   2788.587  5861.621 1975.204  6675.005
Jul 2022      21092.653 13375.893 28809.414 9290.886 32894.420
```

Fig. 10. Forecast of Multiplicative for next 6 months period.

Model 3: Another way is to use software to perform automatic Exponential model selection(ETS). We send our model ts4 to ets in the software find the best fit and give optimum result ets(ts4 , model='ZZZ'), It has taken ETS(M, A, M), where the first letter suggests multiplicative error type, the second letter denotes additive trend type, and the third letter denotes a seasonal type, which is multiplicative.

Results and forecast of the same can be found the below in Fig. 11

```
ETS(M,A,M)

Call:
 ets(y = ts2, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.4095
    beta  = 6e-04
    gamma = 0.5223

  Initial states:
    l = 5566.8327
    b = 246.1577
    s = 0.56 0.3161 0.5049 0.5311 0.7178 1.2421
           0.9664 1.1753 1.4439 1.4481 1.4948 1.5996

  sigma:  0.2184

     AIC     AICc      BIC
6752.396 6754.389 6816.721
> accuracy(ts4)
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -354.3682 2549.813 1540.122 -10.35789 20.12937 0.7050532 0.293539
> forecast(ts4,h=6)
         Point Forecast     Lo 80     Hi 80    Lo 95    Hi 95
Feb 2022      13291.480  9570.754 17012.207 7601.120 18981.840
Mar 2022      12756.300  8890.233 16622.367 6843.661 18668.940
Apr 2022       8716.117  5889.360 11542.873 4392.966 13039.268
May 2022       7029.405  4611.112  9447.699 3330.944 10727.866
Jun 2022       4908.199  3129.135  6687.262 2187.356  7629.042
Jul 2022      23851.513 14791.840 32911.186 9995.939 37707.088
```

Fig. 11. Automatic Exponential model results.

It produces the RMSE value of 2549.813 and AIC of 6752 .389 it makes sense introducing different parameters MAM into the model producing the least RMSE and AIC.

Model 4: Another well-known time series approach is the ARIMA autoregressive integrated moving average, which is well-known for predicting linear functions of recent actual values and residuals and is intended for stationary time series. Because our model has a seasonal component, we chose SARIMA (Seasonal ARIMA), which incorporates an additional seasonal component in the ARIMA model.

(p, d, q) (P, D, Q)m here small p,d,q is a non-seasonal and capital ones are a seasonal part.

Step 1: From the Fig. 1 we believed that our data is not stationary to work on this model we need to do differencing and cross-verify, there are two tests one is the Augmented Dickey-Fuller (ADF) test in R the function is adf.test(ts) where is ts is our model to be verified and if P-value is more than 0.05 it means the model is stationary if not the next method called KPSS the function ndiffs() is used to check what number or order of differentiate to be applied for the model to make it stationary. Our model Fig.12 p-value found to be 0.01 which indicates that our model is already stationary.

```
> adf.test(ts10)

        Augmented Dickey-Fuller Test

data:  ts10
Dickey-Fuller = -9.8634, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Fig. 12. Augmented Dickey-Fuller result.

STEP 2: Next process is to identify the value for the PDQ and pdq this can be done using the ACF and PACF since we have too many values going out of the boundry it is hard to interpret those values so we choose the auto ARIMA plot (auto. arima()) in R to get those values

The optimal value of the model is found to be (1,0,1) (1,1,2) Fig13. here non-seasonal component d is 0 because our model is stationary. The RMSE value was found to be 2247 but in the residual plot ACF lines were moving out of the boundary we tried altering seasonal p and q best RMSE was found for the combination (1,0,1) (3,1,5) Fig. 14

```
> summary(sarima)
Series: ts10
ARIMA(1,0,1)(1,1,2)[12]

Coefficients:
         ar1      ma1    sar1     sma1     sma2
      0.8205  -0.2133  0.6527  -0.8451  -0.0334
s.e.  0.0457   0.0809  0.1292   0.1406   0.0796

sigma^2 = 5330611:  log likelihood = -2868
AIC=5748.01   AICc=5748.28   BIC=5770.49

Training set error measures:
                   ME     RMSE     MAE       MPE    MAPE      MASE       ACF1
Training set 21.46117 2247.616 1315.65 -2.557833 18.5022 0.6022919 0.01119676
```

Fig. 13. Summary of Auto SARIMA.

```
> summary(sarima)
Series: ts10
ARIMA(1,0,1)(3,1,5)[12]

Coefficients:
         ar1      ma1     sar1     sar2     sar3    sma1    sma2     sma3     sma4     sma5
      0.8157  -0.2122  -1.6116  -1.3667  -0.2986  1.5146  1.0780  -0.3842  -0.6848  -0.3351
s.e.  0.0484   0.0859   0.3145   0.3930   0.2987  0.3314  0.4195   0.2659   0.1939   0.1302

sigma^2 = 4773759:  log likelihood = -2863.98
AIC=5749.96   AICc=5750.83   BIC=5791.16

Training set error measures:
                  ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 13.4469 2109.647 1241.117 -2.366025 18.24387 0.5681716 0.0150731
```

Fig. 14. Summary of SARIMA model.

The ACF and residuals graphs are plotted to cross-verify the model. Fig 15 shows three lines crossing the line of null hypothesis threshold of correlation between the series and lag. When there is no connection between the series and lag, the series is deemed white noise, alerting us that our model has white noise it suggests improvements could be made to the predictive model.[2]

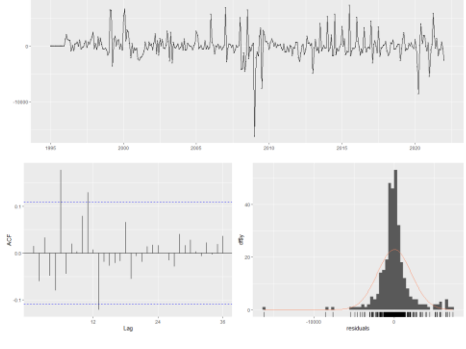

Fig. 15. Residuals Plot for (1,0,1) (3,1,5) combination.

Box-Ljung test: The autocorrelations are not substantially different from zero, according to the test, p-value found to be 0.78. which suffice the requirement and the forecast and plot for the best model is shown in Fig. 16

```
> forecast(sarima,h=6)
         Point Forecast      Lo 80      Hi 80      Lo 95    Hi 95
Feb 2022      10413.796  7531.6030 13295.989  6005.8622 14821.73
Mar 2022       8459.463  5093.1101 11825.816  3311.0703 13607.86
Apr 2022       7355.953  3702.8489 11009.058  1769.0122 12942.89
May 2022       6179.938  2347.9175 10011.958   319.3684 12040.51
Jun 2022       4058.036   111.4685  8004.604 -1977.7183 10093.79
Jul 2022      18741.480 14720.5092 22762.451 12591.9357 24891.03
```

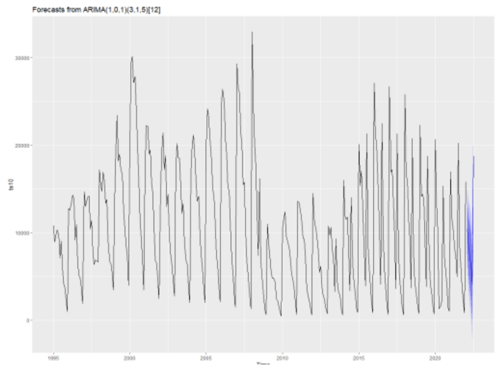

Fig. 16. Next 6-period forecast and its plot.

D. *Results and Interpretation:*

Seasonal Naive was the first model we chose. Our model exhibits a seasonal tendency, as seen in Figures 1 and 2.

Exponential smoothing model 2 We used a triple exponential model, often known as Holt-winters exponential smoothing, because our model has a level, trend, and seasonal (alpha, beta, and gamma) component.

Model 3: Another technique to forecast is to utilize software to select an exponential model automatically.

Finally, Model 4 SARIMA . We picked SARIMA (Seasonal ARIMA), which integrates an additional seasonal component in the ARIMA model, because our model has a seasonal component.

Methods	**RMSE**	**AIC**
Seasonal Naïve	3475.13	-
Holt-winters (Additive)	2604	7026
Holt-winters (Multiplicative)	2291.2	6815.1
Automatic Exponential model	2549.813	6752.3
Auto SARIMA	2247.616	5748
SARIMA	2109	5750

Fig. 17. Summary Table of all results.

Fig. 17 The least RMSE value was discovered in Auto SARIMA and AIC in SARIMA, despite the fact that these models produce superior results, they are more complex and have extreme lines in residual ACF plot, indicating that the model still has to be improved. So, with a value of 2291.2, we chose the simple next best and least RMSE Holt-Winters(Multiplicative) as our best model.

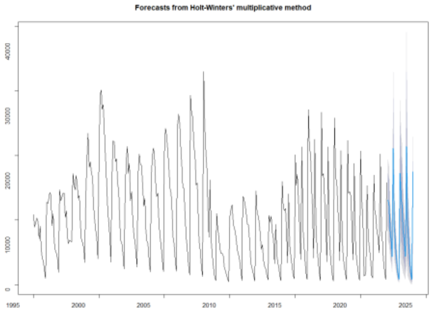

Fig. 18. Multiplicative method plot (Holt-Winters).

Interpretation: Forecast values for the next 6 months can be seen in Fig 18. the blue coloured line shows the forecasted value, the black one shows the original time series, coming to the prediction interval region shaded grey area gives a 95% confidence interval and the blue gives an 80% confidence interval.

II. BINARY LOGISTIC REGRESSION

The goal of this study is to use Binary logistic regression to identify the binary outcome using various predictor factors. All the assumptions are considered to create the model and verified for the excellent fit

A. *Description of the Dataset, Analysis, and Assumptions:*

The data contains 10 Columns, 2722 Rows, and the target column is Default is encoded with 0 and 1, 0 means the person has no default in history and 1 is the person has a default on record, based on the person's characteristics predictor columns model will be built with the best predictors and compared to actual column to check for accuracy

Name	**Description**
Gender	Gender of customer 0-Male,1-Female
Age	Age of the customer
Ed	Years of education
Retire	Retired ,0=not retired, 1=retired
Income	Household income in
Creddebt	Credit card debt in thousands
Othdebt	Other debt in thousands
Marital	Marital status 0=unmarried, 1=married
Homeown	Home ownership,0=rents, 1=owns home
Default	**No default on record (0) / Default on record (1)**

Fig. 19. Data Description.

Block 0: Beginning Block

Classification Table[a,b]

	Observed		Predicted default 0	Predicted default 1	Percentage Correct
Step 0	default	0	1551	0	100.0
		1	1170	0	.0
	Overall Percentage				57.0

a. Constant is included in the model.

b. The cut value is .500

Fig. 20. Null Model.

Fig 20 shows a Null model, which is a baseline for comparison without variables, and it indicates that our prediction model should be greater than 57 percent, also the outcome default looks to be Slightly imbalanced (1151:1170) as we have a less number of rows, we don't split the model to train and test so there will less chance of under and overfitting and we assume oversampling is not necessary.

Parameters to check the goodness of fit:

1. **Omnibus Test:** If the value shows significance, then the model is considered to be fit**.**
2. ***Hosmer and Lemeshow Test:*** The goodness of fit gauged in terms of significance if $p > 0.05$ then the model is considered to be fit, basically alert us if there is a difference between predictor and observed model.
3. ***Nagelkerke R Square:*** It ranges from 0 to 1 close to 1 is considered to be a good fit, it gives the approximate variance in the variables
4. ***Deviance:*** It is the sum of squared residuals, the lower the value is considered to be the better the fit.

B. Steps of model building:

To proceed to build the model it is necessary to analyze the correlation matrix to choose the which variables are affecting the independent variable and check if there is any multicollinearity among the predictor variables

Correlations

		gender	age	ed	retire	income	creddebt	othdebt	default	marital	homeown
gender	Pearson Correlation	1	-.003	.015	.012	-.037	-.028	-.024	-.004	.021	.015
	Sig. (2-tailed)		.892	.444	.537	.054	.138	.205	.851	.267	.432
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
age	Pearson Correlation	-.003	1	-.068**	.556**	.233**	.149**	.160**	-.466**	.016	.016
	Sig. (2-tailed)	.892		.000	.000	.000	.000	.000	.000	.391	.409
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
ed	Pearson Correlation	.015	-.068**	1	-.101**	.202**	.117**	.163**	.121**	-.015	.069**
	Sig. (2-tailed)	.444	.000		.000	.000	.000	.000	.000	.443	.000
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
retire	Pearson Correlation	.012	.556**	-.101**	1	-.173**	-.113**	-.136**	-.276**	.008	-.063**
	Sig. (2-tailed)	.537	.000	.000		.000	.000	.000	.000	.692	.001
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
income	Pearson Correlation	-.037	.233**	.202**	-.173**	1	.728**	.778**	.006	.011	.125**
	Sig. (2-tailed)	.054	.000	.000	.000		.000	.000	.769	.583	.000
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
creddebt	Pearson Correlation	-.028	.149**	.117**	-.113**	.728**	1	.708**	.207**	-.003	.081**
	Sig. (2-tailed)	.138	.000	.000	.000	.000		.000	.000	.875	.000
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
othdebt	Pearson Correlation	-.024	.160**	.163**	-.136**	.778**	.708**	1	.128**	-.003	.084**
	Sig. (2-tailed)	.205	.000	.000	.000	.000	.000		.000	.856	.000
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
default	Pearson Correlation	-.004	-.466**	.121**	-.276**	.006	.207**	.128**	1	-.032	-.050**
	Sig. (2-tailed)	.851	.000	.000	.000	.769	.000	.000		.095	.010
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
marital	Pearson Correlation	.021	.016	-.015	.008	.011	-.003	-.003	-.032	1	.138**
	Sig. (2-tailed)	.267	.391	.443	.692	.583	.875	.856	.095		.000
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721
homeown	Pearson Correlation	.015	.016	.069**	-.063**	.125**	.081**	.084**	-.050**	.138**	1
	Sig. (2-tailed)	.432	.409	.000	.001	.000	.000	.000	.010	.000	
	N	2721	2721	2721	2721	2721	2721	2721	2721	2721	2721

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 21. Correlation Matrix.

In the Fig. 21 matrix, the notable observation multi correlation can be found in the Income, creddebt, and other debt with a value of more than 0.7, and gender income and marital are the column least correlating with the target variable.

Model 1: In the first model all the predictor variables are taken into the consideration with the default threshold value set of 0.5, omnibus, Hosmer-Lemeshow Test, and Nagelkerke R Square values found to be 0.0(significant), 0.39, and 0.49 respectfully although the model pass most of the assumptions in the wald's test few variables like gender, retire and martial were found to be the insignificant result of the same are captured in Fig. 22 and 23.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1246.663	9	.000
	Block	1246.663	9	.000
	Model	1246.663	9	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2471.920[a]	.368	.493

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	8.397	8	.396

Fig. 22. Model 1 results.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender	-.023	.100	.052	1	.820	.978
	age	-.087	.005	353.343	1	.000	.916
	ed	.082	.016	25.312	1	.000	1.085
	retire	-.063	.331	.037	1	.848	.939
	income	-.020	.002	78.801	1	.000	.980
	creddebt	.492	.034	214.108	1	.000	1.635
	othdebt	.115	.017	46.099	1	.000	1.122
	marital	-.020	.100	.039	1	.843	.980
	homeown	-.355	.105	11.486	1	.001	.701
	Constant	1.916	.286	44.949	1	.000	6.793

a. Variable(s) entered on step 1: gender, age, ed, retire, income, creddebt, othdebt, marital, homeown.

Fig. 23. Variables selected, Walds, and odd ratio result.

Model 2: Gender and marital were shown to be less insignificant with the target variable in the correlation matrix, therefore the same shown insignificance in the wald along with retire are removed in this model, The values for the omnibus, Hosmer-Lemeshow Test, and Nagelkerke R Square were 0.0 (significant), 0.50, and 0.49, respectively. We also calculated the -2 log-likelihood value, which was 2472.057, and the model's accuracy was 78.2 percent, with all of the values in the wald test being significant Fig 24, 25.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1246.526	6	.000
	Block	1246.526	6	.000
	Model	1246.526	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2472.057[a]	.368	.493

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	7.328	8	.502

Fig. 24. Model 2 results.

Classification Table[a]

Observed			Predicted: default 0	Predicted: default 1	Percentage Correct
Step 1	default	0	1223	328	78.9
		1	266	904	77.3
	Overall Percentage				78.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.088	.004	467.867	1	.000	.916
	ed	.082	.016	25.336	1	.000	1.085
	income	-.020	.002	80.234	1	.000	.981
	creddebt	.492	.034	215.691	1	.000	1.636
	othdebt	.115	.017	46.212	1	.000	1.122
	homeown	-.358	.104	11.918	1	.001	.699
	Constant	1.907	.274	48.273	1	.000	6.734

a. Variable(s) entered on step 1: age, ed, income, creddebt, othdebt, homeown.

Fig. 25. Confusion matrix and Variables summary.

Model 3 and 4: Even though the model's early assumptions were met, when the variables were evaluated for multicollinearity, we discovered that other debt, credit, and income all had values more than 0.7. The goal of this model is to improve accuracy and lower the -2log probability value. Creddebt is dropped and model preliminary assumptions were found to be satisfied Fig. 26 and gave the accuracy of 0.4 less than the previous model and -2 log value deviance increased to 2519.5, but this time VIF of the model has all less than 5 which satisfies the multicollinearity model 3 considered to be better for next evaluation

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1199.009	5	.000
	Block	1199.009	5	.000
	Model	1199.009	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2519.574[a]	.356	.478

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	6.938	8	.543

Fig. 26. Model 3 results.

Two of our variables had skewness and outliers as we have a very less number of rows to predict using the box plot Fig. 27 we removed only potential outliers

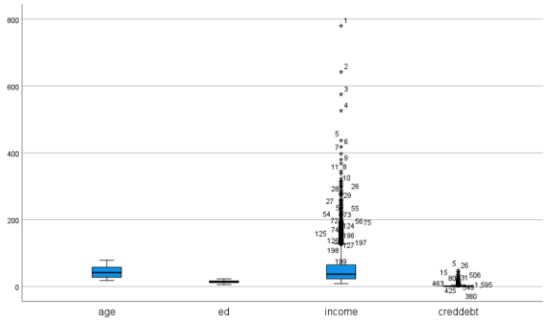

Fig. 27. Box plot to check outliers.

Another crucial assumption to make before deciding on model 3 as the best is that each of the continuous predictive variables is linearly related to the log of the outcome variable, which is known as logit linearity. To check this, we must make sure that the interaction between the continuous variable and the log itself; otherwise, the linearity of the logit assumption will be violated. The continuous variable in the present model are age, ed, income, and credebt Fig. 28 shows the interaction between each other

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.104	.096	1.164	1	.281	.901
	ed	-.238	.442	.291	1	.589	.788
	income	-.089	.014	38.054	1	.000	.914
	creddebt	1.095	.097	128.228	1	.000	2.989
	homeown	-.384	.106	13.178	1	.000	.681
	Log_AGE by age	.003	.020	.026	1	.871	1.003
	Log_ed by ed	.089	.120	.548	1	.459	1.093
	Log_Income by income	.013	.002	29.593	1	.000	1.013
	Log_creddept by creddebt	-.224	.034	42.827	1	.000	.799
	Constant	3.734	1.935	3.721	1	.054	41.827

a. Variable(s) entered on step 1: age, ed, income, creddebt, homeown, Log_AGE * age , Log_ed * ed , Log_Income * income , Log_creddept * creddebt .

Fig. 28. Variables summary after applying LOGIT.

It is found that income and creadebt both are showing the significance sig. a value less than 0.05 which indicates the presence of linearity of the logit as we have a very less number of variables to predict in order to satisfy the non-linearity, one way is to execute transformations by including higher-order polynomial components[3].

Because the majority of the values are 0-1 we use square root [4] to rescale the variables. The procedure is repeated, and now the income and creadebt Fig. 29 shows insignificance.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.107	.102	1.115	1	.291	.898
	ed	-.258	.442	.342	1	.559	.772
	income	-.018	.092	.040	1	.842	.982
	homeown	-.390	.106	13.531	1	.000	.677
	SQRT_Income	-.366	.467	.613	1	.434	.694
	SQRT_Creddebt	1.785	.169	111.531	1	.000	5.957
	Log_AGE by age	.004	.021	.035	1	.851	1.004
	Log_ed by ed	.094	.120	.613	1	.434	1.099
	Log_Income by income	.004	.012	.125	1	.723	1.004
	Log_creddept by creddebt	.026	.018	1.986	1	.159	1.026
	Constant	4.039	2.063	3.832	1	.050	56.758

a. Variable(s) entered on step 1: age, ed, income, homeown, SQRT_Income, SQRT_Creddebt, Log_AGE * age , Log_ed * ed , Log_Income * income , Log_creddept * creddebt .

Fig. 29. Variables summary after LOGIT significance.

After that, all predictor variables are run through the model except the log interaction, and the insignificant and multicollinear variables income and creddept are removed while keeping their square roots and run through the final model 4 again Now the model returns a significant omnibus 0.0, an insignificant Hosmer-Lemeshow Test value of 0.562, and an insignificant Nagelkerke R Square of 0.5, with a model accuracy of 79(78.91), which is the best so far for the default threshold of 0.5 and the least -2 log value deviation value of 2425. result of the same shown in Fig 30.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1256.642	5	.000
	Block	1256.642	5	.000
	Model	1256.642	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2425.928[a]	.373	.500

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	6.764	8	.562

Fig. 30. Model 4 results.

Classification Table[a]

Observed			Predicted: default 0	Predicted: default 1	Percentage Correct
Step 1	default	0	1215	316	79.4
		1	253	909	78.2
	Overall Percentage				78.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.090	.004	460.253	1	.000	.914
	ed	.079	.017	22.947	1	.000	1.083
	homeown	-.394	.105	13.943	1	.000	.674
	SQRT_Income	-.257	.031	69.593	1	.000	.773
	SQRT_Creddebt	1.945	.106	336.963	1	.000	6.995
	Constant	1.867	.274	46.556	1	.000	6.468

a. Variable(s) entered on step 1: age, ed, homeown, SQRT_Income, SQRT_Creddebt.

Fig. 31. Confusion matrix and Variables summary of model 4.

C. Results and Interpretation:

To enhance accuracy, lower the -2log probability value, and satisfy all other assumptions, four model operations were performed.

Model 1: was rejected due to the presence of insignificance in wald's test

Model 2: was rejected due to the presence of multicollinearity

Model 3 and 4: By deleting outliers and satisfying the linearity of the logit assumption, Model 3 was made suitable for the next evaluation.

Model 3 was made to be fit for the next evaluation by removing outliers and satisfying the linearity of the logit assumption in Model 4 was cleaned and tested for various thresholds, however, the default threshold of 0.5 was found to be the best.

There are certain other assumptions that must be taken into account in order to adopt Model 4 are
Dependent variable should be categorical: our target column has binary values so it is mutually exclusive
Absence of multicollinearity: No multicollinearity and all VIF found to be less than 5 Fig 32

Coefficients[a]

Model		Collinearity Statistics: Tolerance	Collinearity Statistics: VIF
1	(Constant)		
	age	.918	1.089
	ed	.924	1.082
	homeown	.978	1.023
	SQR_Income	.471	2.125
	Sqrt_creddebt	.517	1.934

Correlation Matrix

	Constant	age	ed	homeown
Constant	1.000	-.367	-.757	-.170
age	-.367	1.000	.023	.076
ed	-.757	.023	1.000	-.065
homeown	-.170	.076	-.065	1.000
SQRT_Income	-.087	-.163	-.279	-.048
SQRT_Creddebt	.095	-.297	.133	-.052

Fig. 32. Left VIF summary and Right Correlation matrix.

Absence of outliers: In Fig. 33, the residual statistics cooks distance value is less than 4/n, indicating that there is no outlier in the model.

Logit's linearity: Using the Box-Tidwell test, all continuous variable interactions were found to be insignificant, confirming the hypothesis. The last one is, that the minimum number of rows per predictor required is 20, however, ours has more than 2000rows.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-.46	1.78	.43	.302	2695
Std. Predicted Value	-2.935	4.452	.000	1.000	2695
Standard Error of Predicted Value	.010	.084	.018	.005	2695
Adjusted Predicted Value	-.46	1.82	.43	.303	2695
Residual	-.928	1.234	.000	.392	2695
Std. Residual	-2.362	3.142	.000	.999	2695
Stud. Residual	-2.364	3.176	.000	1.000	2695
Deleted Residual	-.930	1.260	.000	.393	2695
Stud. Deleted Residual	-2.366	3.181	.000	1.000	2695
Mahal. Distance	.663	122.718	4.998	4.893	2695
Cook's Distance	.000	.036	.000	.001	2695
Centered Leverage Value	.000	.046	.002	.002	2695

a. Dependent Variable: default

Fig. 33. Outliers summary.

Interpretation of model 4: The table of contingency The observed and expected values are shown in Fig 34. The model expected results are closely predicted for both outcomes, indicating that the The model fits the data well.

Contingency Table for Hosmer and Lemeshow Test

		default = 0 Observed	default = 0 Expected	default = 1 Observed	default = 1 Expected	Total
Step 1	1	264	264.489	5	4.511	269
	2	253	256.241	16	12.759	269
	3	235	240.141	34	28.859	269
	4	209	210.439	60	58.561	269
	5	184	169.693	85	99.307	269
	6	134	133.485	135	135.515	269
	7	110	104.913	159	164.087	269
	8	69	76.489	200	192.511	269
	9	50	51.618	219	217.382	269
	10	23	23.492	249	248.508	272

Fig. 34. Contingency table for Hosmer(Model 4).

Classification Table[a]

Observed			Predicted: default 0	Predicted: default 1	Percentage Correct
Step 1	default	0	1215	316	79.4
		1	253	909	78.2
	Overall Percentage				78.9

a. The cut value is .500

Fig. 35. Confusion matrix (Model 4).

Fig.35 shows the confusion matrix, with overall accuracy greater than block 0, and balanced outcomes of 79.4 and 78.2 for no default and default, respectively. The model successfully predicted 79.4% of persons with no default history and 78.2% of people with default history, which are also known as specificity and sensitivity respectively with overall accuracy of 78.9 percent.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.090	.004	460.253	1	.000	.914
	ed	.079	.017	22.947	1	.000	1.083
	homeown	-.394	.105	13.943	1	.000	.674
	SQRT_Income	-.257	.031	69.593	1	.000	.773
	SQRT_Creddebt	1.945	.106	336.963	1	.000	6.995
	Constant	1.867	.274	46.556	1	.000	6.468

a. Variable(s) entered on step 1: age, ed, homeown, SQRT_Income, SQRT_Creddebt.

Fig. 36. Variable summary

In the table Fig 36, The coefficient of impact on the model negatively or positively for the predictor is shown in column B. For every unit change in the predictor, the chance of the outcome changes by Exp-B, which is the anticipated change in Log odds. We have 3 negative and 2 positive values.

The odds ratio of each row is given in the column EXP(B). If the odds ratio is more than one, the chance of falling into the default group is higher than the chance of falling into the non-defaulters category, and vice versa if the odds ratio is less than one. If one probability is the same.

The odds of default for one unit change in credit debt are 6.9 times higher than individuals who don't fall into this category

D. Dimension reduction technique:

One of the DRT is Principal component analysis(PCA), This aids in the transformation of a bigger number of correlated variables into a small number of uncorrelated variables.

Assumptions need to satisfy to apply PCA are some variables correlation should be more than 0.3 we have it in our dataset, minimum sample required is 10-20 we have more than 2000 plus values, the last assumption is KMO and Bartlett's Test the value should be significant.

Model: kaiser's criterion is followed by giving eigenvalue 1 three combinations were predicted we went ahead with the same combination and performed logistic regression against the Default variable the predicted values did satisfy all the assumptions except Hosmer (0.0 significant) so we plotted screen plot Fig 37 and this time we took 7 factors the factors above the elbow.

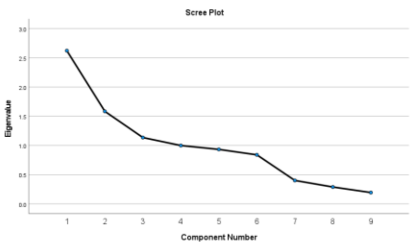

Fig. 37. Scree plot.

Rotated Component Matrix[a]

	Component 1	2	3	4	5	6	7
gender							1.000
age				.918			
ed			.991				
retire		.930					
income	.879						
creddebt	.918						
othdebt	.904						
marital						.997	
homeown					.995		

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Fig. 38. Rotated Component Matrix.

Fig 38 The 7 combinations are shown in a rotated component matrix. Column 1 represents the relationship between income, credit, and other debts, which are dimensionally simplified to make one column.

The new model is run by eliminating columns 6 and 7 because they did not fulfill walds, Result, and discussion: the resultant model met all of the logistic regression assumptions, yielding accuracy and deviance of 77.3 and 2594, respectively. The model predicted 76.1 percent sensitivity and 78.1 percent specificity results of the same shown in Fig 39.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1124.225	5	.000
	Block	1124.225	5	.000
	Model	1124.225	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2594.358[a]	.338	.454

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 1 for analysis 1	1.073	.083	166.966	1	.000	2.923
	REGR factor score 2 for analysis 1	-.190	.073	6.820	1	.009	.827
	REGR factor score 3 for analysis 1	.238	.050	22.940	1	.000	1.269
	REGR factor score 4 for analysis 1	-1.648	.067	612.493	1	.000	.192
	REGR factor score 5 for analysis 1	-.172	.049	12.496	1	.000	.842
	Constant	-.504	.053	90.422	1	.000	.604

Classification Table[a]

Observed			Predicted default 0	1	Percentage Correct
Step 1	default	0	1212	339	78.1
		1	280	890	76.1
	Overall Percentage				77.3

Fig. 39. PCA Model Result summary.

III. CONCLUSION AND FUTURE WORK

With the awareness that our dataset contained trend, seasonality, and some irregularity, we used many approaches to anticipate the following six months period value, including seasonal Naïve, Holt-winters multiplicative, Automatic Exponential model, Auto SARIMA, and SARIMA, Although these SARIMA produce better results, they are more complex, and the residual plot suggested that the model still needs to be improved. We chose the basic next best and least RMSE Holt-Winters (Multiplicative) model with a value of 2291.2.

Logestic regression the final model 4 demonstrated the highest accuracy of 78.9% and the least deviation of 2425 for determining whether the client has default history or not. and finally, the technique of dimension reduction PCA is used, it can help turn a large number of correlated variables into a small number of uncorrelated variables. It was implemented to see if we might improve accuracy, however the overall accuracy was found to be 78.1, which is lower than earlier models.

Time series data is real-world data that would have been satisfied if cleaned and processed before applying the SARIMA model, and more cleaning and including different variables and scaling them would have given the fruitful result in logistic regression.

IV. REFERENCES

[1] Brownlee, "How to Decompose Time Series Data into Trend and Seasonality", Machine Learning Mastery, 2022. [Online]. Available: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/.

[2] Brownlee, "White Noise Time Series with Python", Machine Learning Mastery, 2022. [Online]. Available: https://machinelearningmastery.com/white-noise-time-series-python/.

[3] "Assumptions of Logistic Regression, Clearly Explained", Medium, 2022. [Online]. Available:https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290#:~:text=Logistic%20regression%20does%20not%20require,but%20not%20for%20logistic%20regression.

[4] "Factoring higher degree polynomials (video) | Khan Academy", Khan Academy, 2022. [Online]. Available: https://www.khanacademy.org/math/algebra2/x2ec2f6f830c9fb89:poly-factor/x2ec2f6f830c9fb89:factor-high-deg/v/factor-high-deg-poly.