Application of Time Series Forecasting and Logistic Regression for Binary Classification

Abhinav Thapa
Student ID: 20259409
Statistics for Data Analytics, MSc in Data Analytics
National College of Ireland, Dublin
Email: x20259409@student.ncirl.ie

***Abstract*— This report aims to demonstrate the application of two of the most widely used statistical techniques – Time Series Analysis and Logistic Regression. Time Series Analysis allows to grasp the effects of trend, seasonality and noise on the concerned response variable and build an accurate time-dependent forecast. On the other hand, Logistic Regression is typically used to classify a dichotomous response variable regressed upon a combination of categorical and/or numerical predictors. Both the use-cases are diagnosed and optimized for performance and fit using expansive descriptive statistics, thorough model diagnostics and relevant performance metrics.**

Keywords*— *EDA, Time Series Analysis, Trend, Seasonality, Noise Forecast, Logistic Regression, Feature Selection, Model Diagnostics, ARIMA/ SARIMA, Time series (TS), Correlation, Predictors, Response. Tools used- R, SPSS.

I. INTRODUCTION

The first dataset identified for Time Series Analysis, **CarRegistrations.csv**, contains monthly data from Ireland explained in 2 features – Time in months from Jan 1995 to Jan 2022 and Number of cars registered, and a total of 325 records. We will forecast the number of new private car registrations using various Time Series Analysis methods on the dataset and arrive at an optimum final model that can forecast upto 6 periods ahead. Concurrently, on the second dataset, **Default.csv**, that has 2721 records, we will apply Logistic Regression to determine whether a customer is likely to have a default on their record using necessary model building and diagnostic steps. This dataset contains 5 numerical predictors - Age, Years of education, Income, Credit card debt and other debt, and 4 categorical predictors – Gender, Retired, Marital Status and Home-ownership, apart from the binary response variable – Default.

II. TIME SERIES ANALYSIS

A. *Descriptive Statistics and Preliminary Assessment*

On the monthly timeseries data for new private cars registered from Jan 1995 to Jan 2022, we performed pre-processing, checked for missing values and prepared it for a preliminary assessment in R in order to exhibit its nature.

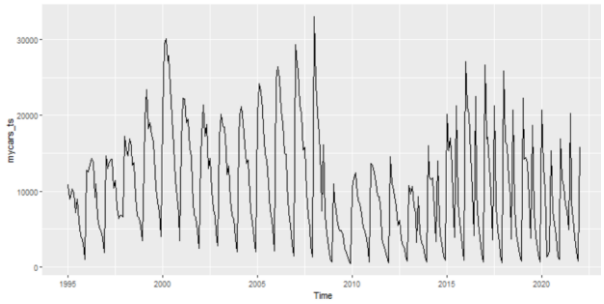


Figure 1: Raw Time Series

Initially, the series did not present as having an additive or multiplicative nature (Fig.1). We observed a lot of seasonality, and an indefinite trend, with some noise. The seasonality effects on the time series were observed in detail in the plots in Fig.2, that also showed how the series had gradual reducing trend throughout the year and that typically in January highest number of new cars get registered while it was the least in December.

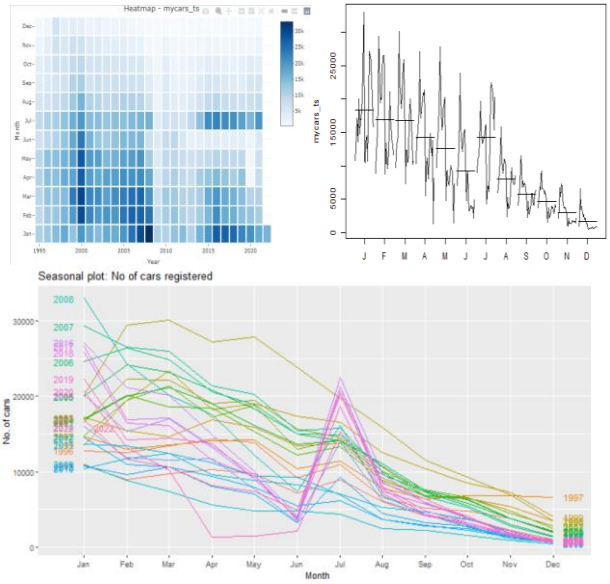


Figure 2: Seasonal nature explained by Heatmap, Monthplot and Seasonal plot.

Further decomposition of the time series revealed that the trend is irregular and seasonality is additive in nature. Since, there not much changes in both the decompositions (additive/ multiplicative), we decided to go with the multiplicative model as it was able to capture more of the patterns from seasonality and random aspects of the series (Fig. 4, the scale of the seasonal and random reduces in multiplicative model).

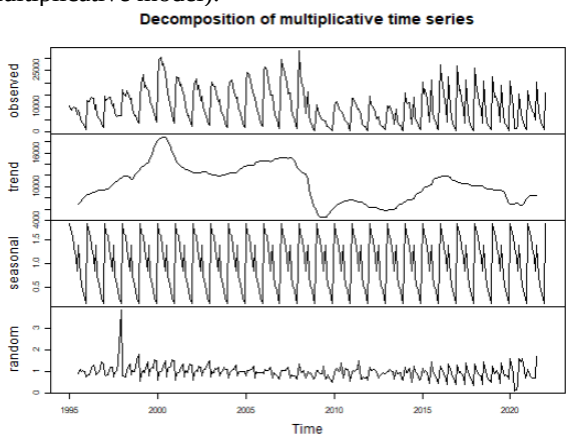


Figure 3: Multiplicative decomposition of Time Series

We modelled several simple models such as **Mean, Naïve** and **Seasonal Naïve models (Fig.4)** and found the respective **RMSEs - 7164.51, 6654.07**, and **3475.13** in order to gauge the generic time series behavior.

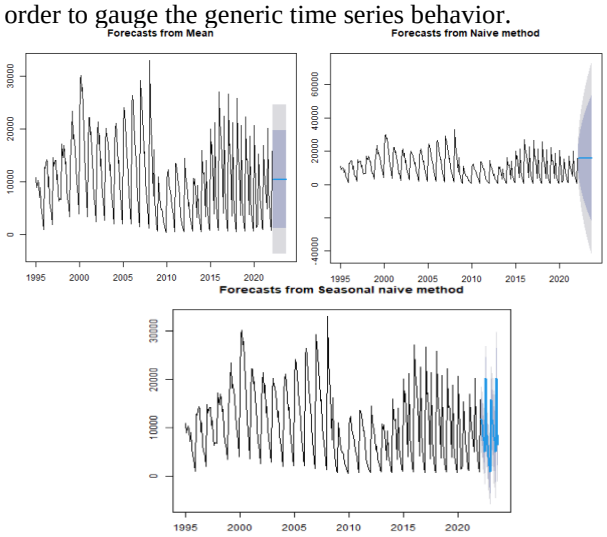


Figure 4: Mean, Naive and S-naive TS models

Now that we have a fair insight into the nature (patterns and irregularities) of the time series, we can proceed to the Model building step where we shall look into more complex and preferably better models for forecast with better accuracy.

B. *Model-Building, Diagnostics and Evaluation*

In time series modelling, arriving at a proper model involves exlporing multiple modelling techniques while checking for accuracy, autocorrelations and the nature of residuals simulateneously. We looked into the RMSE values, Normality plots (Q-Q plots), ACF and PACF plots, and the residuals plots for each model to justify the model building and selection steps taken. So it is a good practice to start with simple models and increase complexity depending upon the problem statement.

1. *Exponential Smoothing Models*

Exponential Smoothing is a technique in which the effects of individual components (trend/ seasonality/ noise) of the TS is smoothed (reduced or manipulated) to gauge the underlying true pattern of the series to forecast accurately.

i. *Simple Exponential Smoothing (SES)*

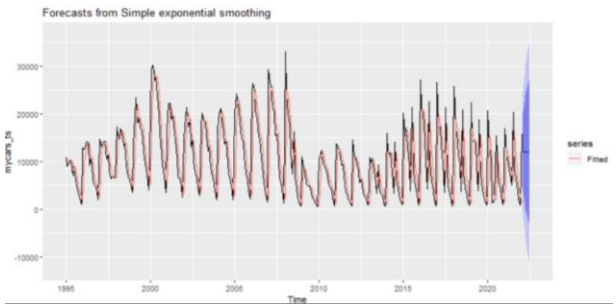


Figure 5: Simple Exponentially Smoothed TS

Here, we started with the simple exponential smoothing technique that 'smoothes' the effects of level by putting more weightage exponentially on the recent lags than on the older ones, and then forecasts.

```
> mycars_ses
         Point Forecast        Lo 80    Hi 80      Lo 95    Hi 95
Feb 2022       12059.64  3730.5918438 20388.68  -678.539 24797.82
Mar 2022       12059.64  1696.5633978 22422.71 -3789.317 27908.59
Apr 2022       12059.64     0.8809855 24118.40 -6382.640 30501.92
May 2022       12059.64 -1484.1401426 25603.42 -8653.783 32773.06
Jun 2022       12059.64 -2821.7003840 26940.98 -10699.405 34818.68
Jul 2022       12059.64 -4048.5753671 28167.85 -12575.749 36695.03
> round(accuracy(mycars_ses),2)
                ME    RMSE     MAE    MPE  MAPE MASE ACF1
Training set 6.99 6479.16 4262.62 -37.67 60.45 1.95 0.03
> Box.test(mycars_ses$residuals, type='Ljung-Box') # signi so problem with
 acf all zero assumpt

        Box-Ljung test

data:  mycars_ses$residuals
X-squared = 0.27065, df = 1, p-value = 0.6029
```

Figure 6: SES Forecast, Accuracy and Box-Ljung Test

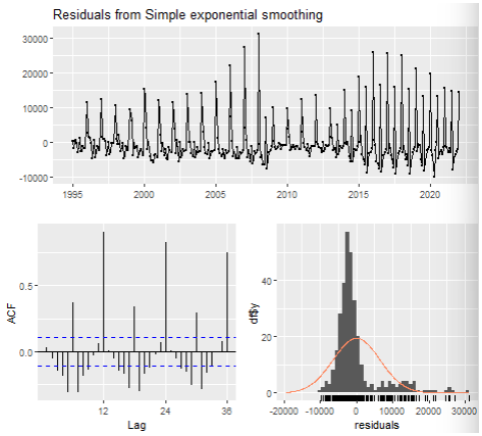


Figure 7: Residuals diagnostics plot

Fig. 5, 6 and 7 reveal **high RMSE, non-normal residuals distribution and high autocorrelations with the lags**, so we shall proceed onto the next model i.e., holt's (double exponential smoothing) model.

ii. *Holt's Linear Trend Method*

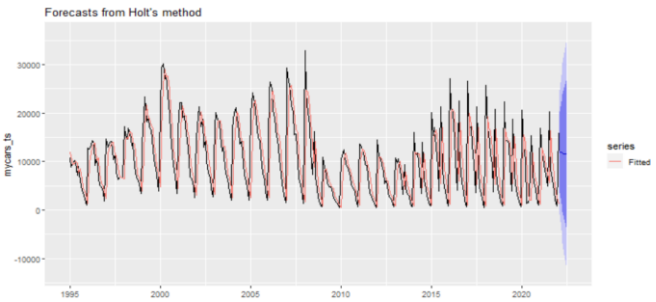
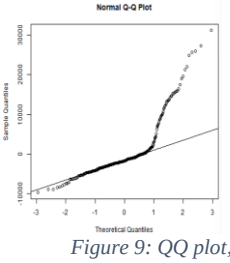


Figure 8: Holt's linear trend model TS with fitted values

This method performs both trend-smoothing and level-smoothing on the TS and models the forecast (Fig.8).




```
> Box.test(mycars_holt$residuals, type='Ljung-Box')

        Box-Ljung test

data:  mycars_holt$residuals
X-squared = 0.25914, df = 1, p-value = 0.6107

> round(accuracy(mycars_holt),2)
                ME    RMSE     MAE    MPE  MAPE MASE ACF1
Training set 126.99 6481.56 4204.8 -34.86 58.05 1.92 0.03
```

Figure 9: QQ plot, Accuracy and Ljung-Box Test

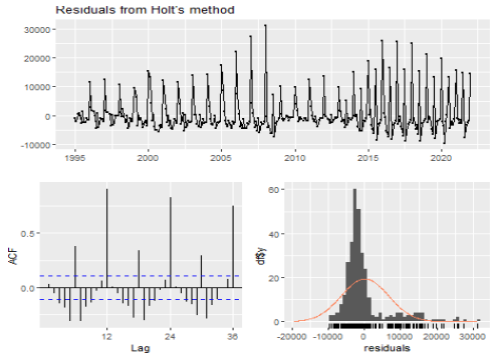


Figure 10: Residuals diag. plots for Holt's model

The **same problems found in *SES* model (i). still persists in this model**. Although a **damped** trend version could be used, a lack of evidence that it could fix all the problems suggests us to move to Holt-Winters model that may be able to correct them.

iii. Holt-Winters Seasonal Method

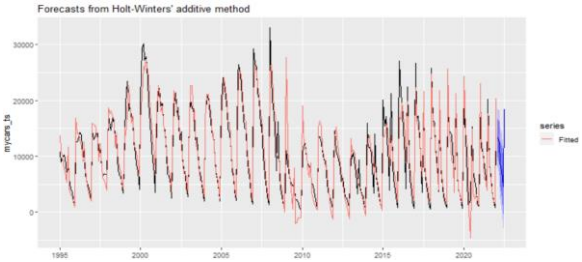


Figure 11: Holt-Winters Model TS

This method performs smoothing on level, trend and seasonality simultaneously for capturing the patterns and then forecasts (default: additive type smoothing).

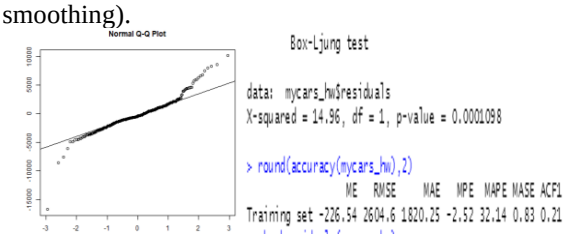


Figure 12: Holt-Winters model QQplot, Diag. test and Accuracy

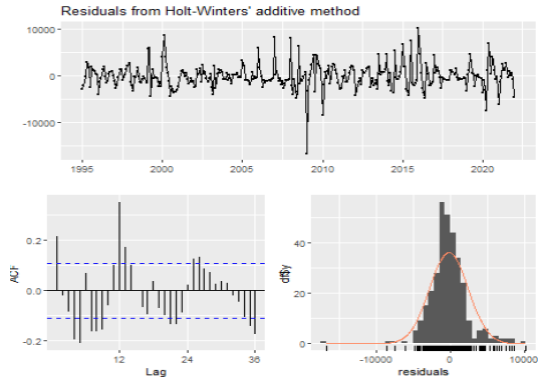


Figure 13: Holt-Winters model diagnostics

Even though the **RMSE** has improved to **2604.6**, the residuals are more normal and the autocorrelations reduced when compared to holt's model, **we can still improve the performance by applying multiplicative type seasonality smoothing** on the time series. The results in Fig. 14 show that the multiplicative model shows better resemblance to the original time series and is more accurate.

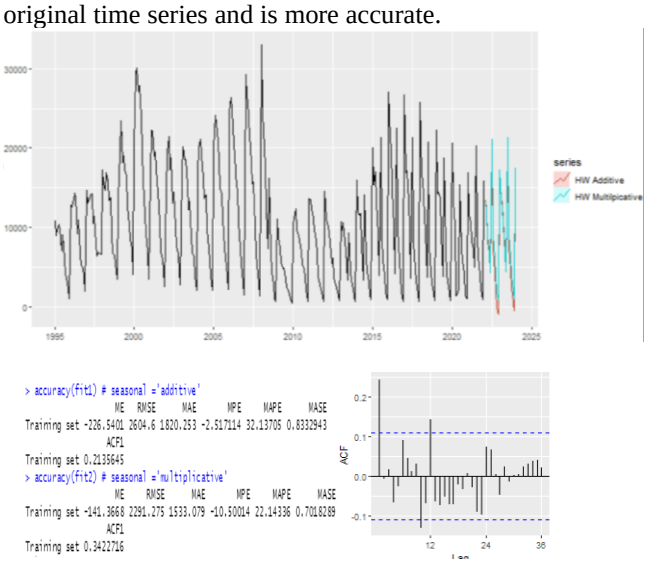


Figure 14: Holt-Winters Additive and Multiplicative Seasonally smoothed models; ACF plot for multiplicative model; and Accuracies

RMSE for multiplicative model (2291.27) is lower than that for additive model. On top of that residual's diagnostics plot is improved and the autocorrelations (ACF plot) have become less erratic.

It is evident that controlling the smoothing parameters and the nature of the smoothing, we can build a model that can perform better as well as provide forecasts free of autocorrelations with its prior lags and with the residuals normal. In order to configure the models at better resolution an ***ets()* - Error Trend Seasonality (ETS)** function can be used for modelling at controlled smoothing settings to build custom models that may perform better than our prior models. We explored the ETS model with "**MMM**" setting and then "**ZZZ**" setting to gauge a better model both manually and automatically to conclude that an automated ETS model with "**MAM**" i.e., **Multiplicatively smoothed at levels and seasonality while multiplicatively smoothed at trend gives the better BIC, however, the manually selected "MMM" model gives the better AIC, AICc and RMSE scores when compared but the diagnostic plots don't change much** (Fig.15). Nevertheless, both ETS models still **don't pass the Box-Ljung Test** that proves that there are still some residual autocorrelations that are not closer to zero and may have an effect on the model.

```
> Box.test(mycars_ets$residuals, type='Ljung-Box') #Here P value should be in-significant

        Box-Ljung test

data:  mycars_ets$residuals
X-squared = 26.341, df = 1, p-value = 2.862e-07

> Box.test(mycars_etsZ$residuals, type='Ljung-Box') #Here P value should be in-significant

        Box-Ljung test

data:  mycars_etsZ$residuals
X-squared = 15.606, df = 1, p-value = 7.8e-05
```

Figure 15: Residual diagnostics for ETS("MAM") model

Since, the earlier models lack the level of complexity and control over the specific elements of the time series, they are unable to achieve the performance and clarity that we desire. So we employed some of the advanced techniques from ARIMA family now.

2. *Autoregressive Integrated Moving Average (ARIMA) Models*

ARIMA models provide control for both the autoregressive terms (lags of stationary series- *p*) and moving average terms (lags of errors- *q*) on a stationary series (differenced- *d*) and forecasts by capturing maximum patterns from the TS.

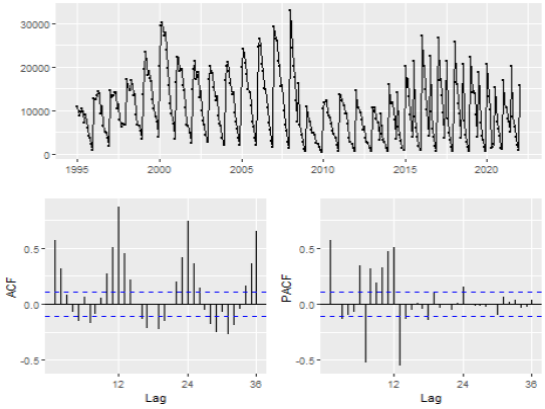


Figure 16: Original Time series with ACF and PACF plots

i. ARIMA Model (Basic)

Fig. 16 showed that we need to fix the variance by normalizing it using Box-Cox transformation (log) on our time series before modelling (Fig.17). We then applied differencing of order 1 as suggested by the *ndiffs()* function.

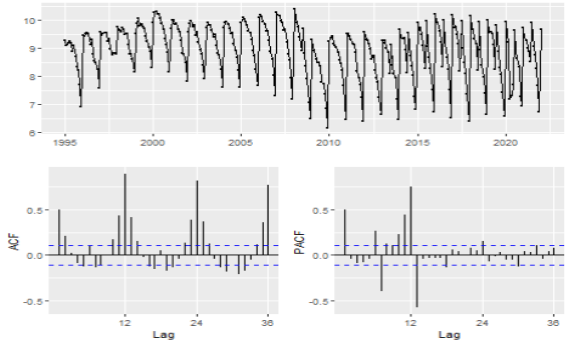


Figure 17: Transformed TS with normalized variance

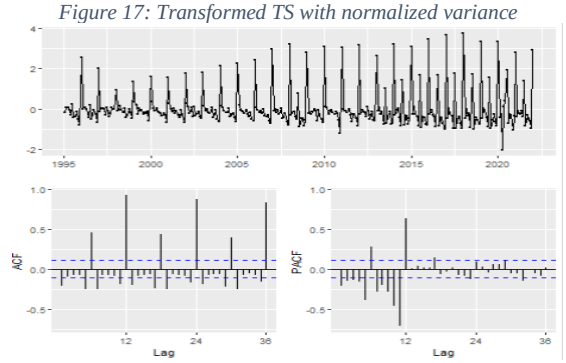


Figure 18: Differenced Fig.17 model

We then checked the stationarity using the Dickey-fuller test and found that the time series is stationary in nature and proceeded with the ARIMA model building. Observing the corresponding ACF and PACF plots in Fig. 18 carefully, we observed that among the earlier lags abrupt patterns suggest possible values of AR terms (*p*) as (1/2/4) and MA(*q*) as (1/2/3) Therefore, we will model using combinations of these *p* and *q* (with *d*=1) terms to arrive at our final best ARIMA model.

Model building in multiple iterations for ARIMA resulted in models with orders **(*p,d,q*)**:

1. (1,1,1): AIC: 6553.37; RMSE: 5887.43; Residuals distribution: skewed positively
2. (1,1,2): AIC: 6555.06; RMSE: 5884.82; Residuals distribution: skewed positively
3. **(1,1,3): AIC: 6448.04; RMSE: 4951.23; Residuals distribution: Normal**
4. (4,1,2): AIC: 6495.4; RMSE: 5299.01; Residuals distribution: slightly skewed positively
5. (4,1,3): AIC: 6472.8; RMSE: 5099.45; Residuals distribution: skewed positively
5. (1,0,3): AIC: 6561.78; RMSE: 5747.8; Residuals distribution: skewed positively

Although **all the above models passed the Box-Ljung test**, but the best model turned out to be model no.3 (1,1,3).

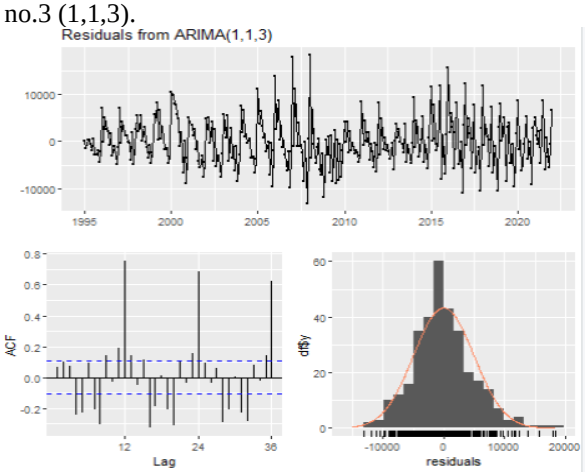
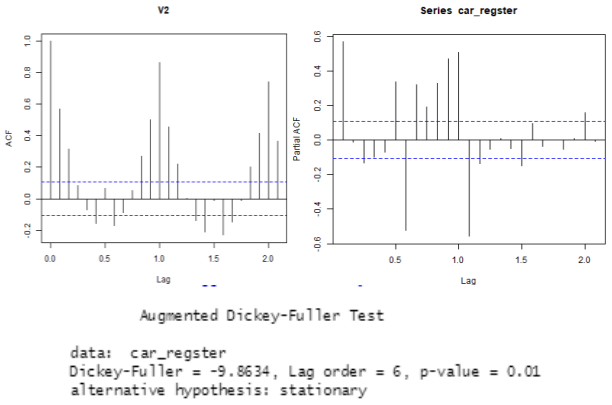


Figure 19: ARIMA Model #3 Residuals plot

Nevertheless, **we can still see some seasonal patterns in the residuals (Fig. 19) and many significant spikes in the ACF which indicate that handling these seasonal patterns may improve our model for accuracy and fit**. In the next step we shall be delving into SARIMA models for the same.

ii. Seasonal- ARIMA (SARIMA) Model

Observing the ACF and PACF plots we can arrive at the non-seasonal and seasonal i.e., (p,d,q) and (P,D,Q) terms for the SARIMA model building that uses these terms to damp the effects of significant seasonality on the time series and helps capture underlying true patterns of time series. We followed similar initial steps as done in ARIMA modelling (Fig.20).




```
        Augmented Dickey-Fuller Test

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

Figure 20: ACF, PACF plots for original TS and Stationarity check.

Then we started modelling with the SARIMA model with different combinations of seasonality terms (abruptness of PACF and ACF spikes being the inspiration for *p*, *q*, *P* and *Q* terms) and checked each model for performance metrics to arrive at the final SARIMA model. Following results were obtained:

Model building in multiple iterations for SARIMA resulted in models with orders **(*p, d, q*), (*P, D, Q*)**:

1. (1,0,2), (1,1,2): AIC: 5748.97; RMSE: 2244.07; Residuals distribution: Normal
2. (1,1,2), (1,1,1): AIC: 5735.2; RMSE: 2249.41; Residuals distribution: Normal
3. (1,1,3), (1,1,1): AIC: 5736.02; RMSE: 2245.72; Residuals distribution: Normal
4. **(3,1,4), (1,2,3): AIC: 5581.62; RMSE: 2194.99; Residuals distribution: Normal**
5. (3,2,4), (1,2,3): AIC: 5589.62; RMSE: 2295.1; Residuals distribution: Normal

All the above models also passed the Box-Ljung test, but the best models turned out to be model no.4 **{(3,1,4), (1,2,3)}.**

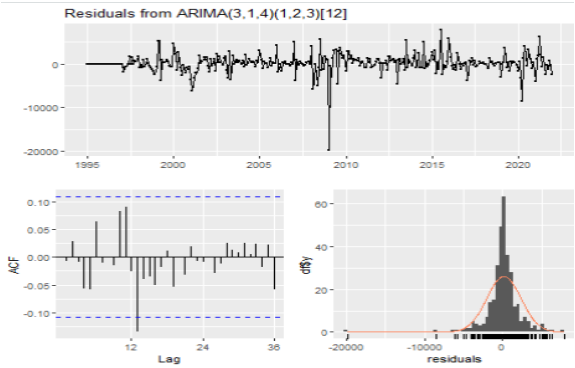


Figure 21: Residual daig. plot for final SARIMA model

Finally, **we can see that the residuals are normal, fairly pattern free and no significant ACF spikes in the recent lags (Fig. 21) which indicate that we have arrived at our final SARIMA model by handling seasonal patterns in ARIMA model and improved our model accuracy and fit**. This concludes time series modelling here.

III. LOGISTIC REGRESSION

A. *Descriptive Statistics and Exploratory Data Analysis*

In order to develop a concrete understanding of the data we need to use descriptive statistics to get the initial overview and then we could strategize for the EDA step. We can get initial description using the following commands in R.

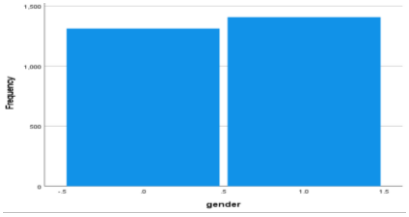
```
> str(mydata)
'data.frame':   2721 obs. of  10 variables:
 $ gender  : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 1 1 ...
 $ age     : int  75 63 53 61 31 46 47 58 71 25 ...
 $ ed      : int  16 13 15 16 15 10 15 11 8 17 ...
 $ retire  : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 1 ...
 $ income  : int  13 55 36 33 20 54 62 125 146 13 ...
 $ creddebt: num  0.4973 1.3901 0.4186 0.0755 0.2047 ...
 $ othdebt : num  0.829 2.735 0.625 1.376 2.175 ...
 $ default : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ marital : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 2 1 ...
 $ homeown : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...

> summary(mydata)
 gender        age              ed           retire          income
 0:1313   Min.   :18.00   Min.   : 6.00   0:2412   Min.   :   9.00
 1:1408   1st Qu.:28.00   1st Qu.:12.00   1: 309   1st Qu.:  23.00
          Median :42.00   Median :15.00            Median :  37.00
          Mean   :43.91   Mean   :14.76            Mean   :  54.69
          3rd Qu.:58.00   3rd Qu.:17.00            3rd Qu.:  64.00
          Max.   :79.00   Max.   :23.00            Max.   :1073.00
    creddebt          othdebt          default  marital  homeown
 Min.   :  0.00136   Min.   :  0.0167   0:1551   0:1429   0: 995
 1st Qu.:  0.42412   1st Qu.:  1.0539   1:1170   1:1292   1:1726
 Median :  1.00036   Median :  2.1961
 Mean   :  2.20815   Mean   :  3.9295
 3rd Qu.:  2.27236   3rd Qu.:  4.6438
 Max.   :109.07260   Max.   :141.4591
```

Figure 22: Descriptive Statistics in R

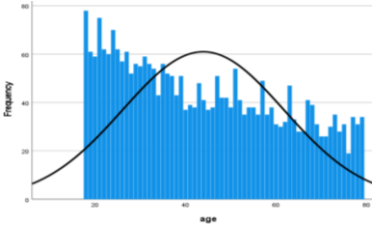
i. *gender*

Gender is a binary categorical feature that represents Male class by '0' and Female by '1' in the dataset.



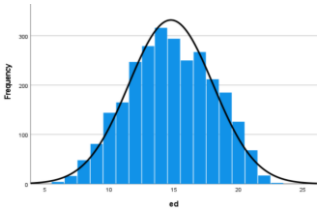
ii. *age*

Age is a continuous variable that represents the age of each of the individuals in years in each record.



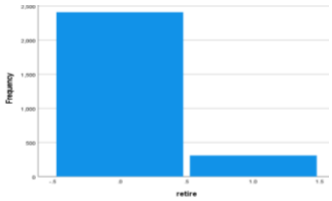
iii. *ed*

This is a continuous feature that represents the years of education that a person has undergone in years.



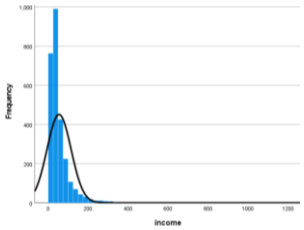
iv. *retire*

It is the categorical feature that represents whether the person has retired or not.



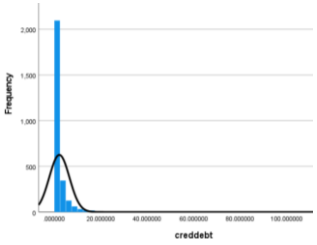
v. income

This feature stands for numerical data regarding the household income in thousands for each person in the dataset.



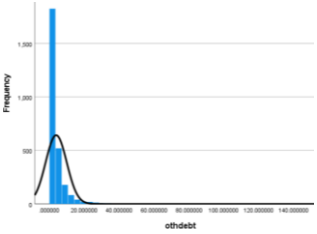
vi. creddebt

It is another continuous feature that gives the credit card debt in each record in thousands.



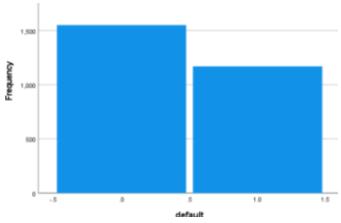
vii. othdebt

This variable represents continuous data for other debt for each record.



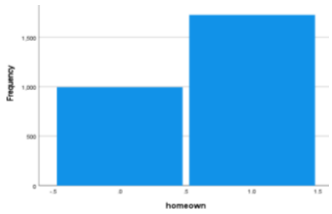
viii. default

This is the dichotomous response variable for our study which we shall use for building our model.



ix. homeown

It is the categorical feature that represents the binary data that whether the person 'owns' or 'rents' his home.



From the initial descriptive statistics in **R** and **SPSS**, we can make the following observations:

i. *age* is fairly positively skewed and not normal in nature, and it may need transformations for model building.

ii. *ed* is normally distributed about the mean.

iii. *income, creddebt and othdebt, are highly positively skewed about their means* which may be due to presence of extreme outliers (Fig.2). It may need some transformations.

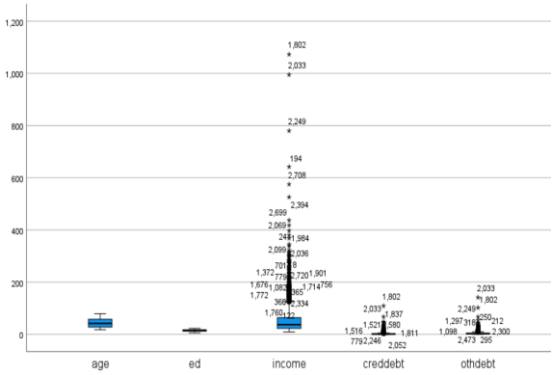


Figure 2: Boxplots for each numerical variable

iv. *gender and default* categorical variables have almost equally distributed binary classes, while *retire* and *homeown* have highly imbalanced binary classes i.e., almost 9 times of population are 'not retired' for every 1 person 'retired' and almost twice as many people 'own' as those who 'rent' their house.

We can also determine the correlation among the numerical variables by using the corelation function and

its corresponding plots using pairs() method in R (Fig.3,4)

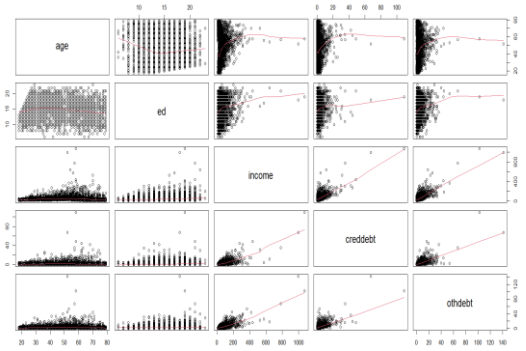


Figure 3: Pearson correlation among numerical variables in the dataset in R.

Correlations

		age	ed	income	creddebt	othdebt
age	Pearson Correlation	1	-.068**	.233**	.149**	.160**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	2721	2721	2721	2721	2721
ed	Pearson Correlation	-.068**	1	.202**	.117**	.163**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	2721	2721	2721	2721	2721
income	Pearson Correlation	.233**	.202**	1	.728**	.778**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	2721	2721	2721	2721	2721
creddebt	Pearson Correlation	.149**	.117**	.728**	1	.708**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	2721	2721	2721	2721	2721
othdebt	Pearson Correlation	.160**	.163**	.778**	.708**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	2721	2721	2721	2721	2721

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

Figure 4: Correlation in SPSS.

Careful evaluation of the correlations provides us with the following observations:

i. *age* is non-linearly correlated with *ed, income, creddebt* and *othdebt.*

ii. *ed* is approximately linearly correlated with *income, creddebt* and *othdebt.*

iii. *income, creddebt and othdebt* are highly positively correlated.

iv. Presence of **Highly correlated predictors (Pearson correlation)** suggest a possibility of Multicollinearity among them that may need to be considered while model building (correlation > 0.25 is considered as high).

With the knowledge of the dataset and the related nature of its features extracted from the initial exploratory data analysis step, we can consider few possible findings before model building, which are as follows:

i. Since the categorical features are all binary in nature, we don't necessarily need to typecast them (into factor types in R) for model building.

ii. Since the dependent variable *default* is fairly equally distributed between '0' and '1' labels, we won't need to treat for class imbalance.

iii. The Dataset contains enough sample size (over 2500 records) and has a dichotomous dependent variable *default* that means that it meets the **requirements for Logistic regression**.

iv. **No missing values** were present in the dataset and the name of the feature *gender* was corrected in R.

```
> linfit2<- lm(mydata$default~.-gender-retire-marital-homeown,data = mydata)
> vif(linfit2)
     age       ed   income creddebt  othdebt
1.075249 1.060970 3.179262 2.393346 2.847087
```

Figure 5: VIF test of Multicollinearity for numerical features.

v. Fig. 5 shows that the numerical predictors satisfy the **assumption (vif scores<5)** that they are not **Multicollinear** in nature.

We shall delve into other assumptions for Logistic Regression such as Linearity of Logit, Influential Outliers, and Independence of errors (Residuals), in the next step. Let's proceed to the data transformations, model building and diagnostic step.

B. *Model Building and Diagnostics*

In model building step, we shall start with the base model (with all initial predictors) and proceed in iterations towards the best logistic regression model that presents best Accuracy, Goodness of Fit and meets the assumptions criteria, using various statistical tests and the rule of parsimony.

i. Iteration 1 (Base Logistic Regression Model)

In this model we included all the initial nine features and used for modelling against *default* variable to see the performance of our *base model* in terms of its relevant metrics **(AIC** and **Deviance)** and to check the contributing variables. Fig. 6 shows the results for the base model in the 1st iteration in R. **AIC is 2491.9 and Deviance is 2471.9**.

```
Call:
glm(formula = default ~ ., family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0136  -0.7030  -0.2147   0.7778   4.0582

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.915863   0.285761   6.704 2.02e-11 ***
gender      -0.022720   0.099653  -0.228 0.819656
age         -0.087255   0.004642 -18.797  < 2e-16 ***
ed           0.081927   0.016284   5.031 4.88e-07 ***
retire      -0.063270   0.330949  -0.191 0.848387
income      -0.019695   0.002219  -8.877  < 2e-16 ***
creddebt     0.491738   0.033606  14.632  < 2e-16 ***
othdebt      0.114857   0.016916   6.790 1.12e-11 ***
marital     -0.019930   0.100480  -0.198 0.842774
homeown     -0.354645   0.104642  -3.389 0.000701 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3718.6  on 2720  degrees of freedom
Residual deviance: 2471.9  on 2711  degrees of freedom
AIC: 2491.9

Number of Fisher Scoring iterations: 6
```

Figure 6: Base model summary

We also checked for the concerned assumptions with the base model and found that there were no problems of Residuals and Influential Outliers in the dataset (Fig. 7) as the standard residuals with absolute values above 2 are less than 5% and above 2.5 are less than 1% and no cook's distance is above 1.

```
> # check residuals and outliers
> sum(rstandard(logreg1)>2.5) # 11 i.e. 0.4 %
[1] 11
> sum(rstandard(logreg1)>2)# 43 i.e. 1.5 %
[1] 43
> sum(cooks.distance(logreg1)>1)
[1] 0
```

Figure7: Residuals and Influential Outlier Assumptions verification

Checking for Linearity of Logit Assumption: We still need to satisfy the linearity between predictors and the logit assumption using Box-Tidwell test in order the proceed with the model building step and we did that by regressing **only the numerical features** and their **corresponding *Interaction variables*** against our response variable *default* and evaluating whether we have any significant interaction terms in the logistic model or not.

```
Call:
glm(formula = default ~ age + age:log(age) + ed + ed:log(ed) +
    income + income:log(income) + creddebt + creddebt:log(creddebt) +
    othdebt + othdebt:log(othdebt), family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4240  -0.6729  -0.2182   0.7431   2.8307

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.2709673  1.9459867   1.681 0.092787 .
age                  -0.0853864  0.0958323  -0.891 0.372930
ed                   -0.2289830  0.4441528  -0.516 0.606169
income               -0.1134887  0.0151411  -7.495 6.61e-14 ***
creddebt              0.8852386  0.1050362   8.428  < 2e-16 ***
othdebt               0.3445217  0.0720583   4.781 1.74e-06 ***
age:log(age)         -0.0003033  0.0201942  -0.015 0.988015
ed:log(ed)            0.0862475  0.1207346   0.714 0.475007
income:log(income)    0.0164693  0.0025653   6.420 1.36e-10 ***
creddebt:log(creddebt) -0.1652033  0.0379264  -4.356 1.33e-05 ***
othdebt:log(othdebt)  -0.0714400  0.0210426  -3.395 0.000686 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3718.6  on 2720  degrees of freedom
Residual deviance: 2428.7  on 2710  degrees of freedom
AIC: 2450.7

Number of Fisher Scoring iterations: 6
```

Figure 8: Summary of model for Box-Tidwell test

It is clear from Fig.8 that the statistically significant interaction terms show that variables – *income, creddebt* and *othdebt* are not linearly related to the log-odds of the outcome variable – *default,* and must be transformed into their logs and only then used in a model. In the next iteration we implemented the proposed action and build another model.

ii. Iteration 2 (log transformed features added)

Here we implemented the logistic model by including the log transformations of the *income, crededbt* and *othdebt* variables and results were obtained (Fig.9) with **AIC 2440.5** and **Residual deviance 2414.5**.

```
glm(formula = default ~ gender + age + ed + retire + income +
    creddebt + othdebt + marital + homeown + log(income) + log(creddebt) +
    log(othdebt), family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4857  -0.6701  -0.2114   0.7216   3.0631

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.851124   0.633069   9.242  < 2e-16 ***
gender       -0.014327   0.101194  -0.142 0.887413
age          -0.082820   0.004973 -16.653  < 2e-16 ***
ed            0.091184   0.016813   5.423 5.85e-08 ***
retire       -0.622947   0.362898  -1.717 0.086054 .
income       -0.002612   0.002806  -0.931 0.351978
creddebt      0.366465   0.041026   8.933  < 2e-16 ***
othdebt       0.044933   0.020910   2.149 0.031645 *
marital       0.015828   0.102219   0.155 0.876948
homeown      -0.373024   0.106622  -3.499 0.000468 ***
log(income)  -1.327938   0.204749  -6.486 8.83e-11 ***
log(creddebt)  0.275717   0.079550   3.466 0.000528 ***
log(othdebt)  0.363116   0.094295   3.851 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3718.6  on 2720  degrees of freedom
Residual deviance: 2414.5  on 2708  degrees of freedom
AIC: 2440.5
```

Figure 9: Iteration 2 Model Summary

It is clear that both the AIC and Residual Deviances have improved in this model compared to the Base model in iteration 1, however, there are still some insignificant features (p>0.05) that can be removed to improve the model (*gender>marital>income>retire*). Also, we cannot negate the possible effects of the Interaction variables on the model which will be explored in iteration 3.

```
> # check residuals and outliers
> sum(rstandard(logreg1_1)>2.5) # 7 i.e. 0.2 % < 1%
[1] 7
> sum(rstandard(logreg1_1)>2)# 46 i.e. 1.6 % < 5 %
[1] 46
> sum(cooks.distance(logreg1_1)>1)
[1] 0
```

Figure 10: Checking for Influential Outliers, Irregular data points

And it is evident from Fig. 10, there is no sign of violation of the assumptions for logistic regression as well.

iii. *Iteration 3 (Interaction terms added)*

We explored all combinations of the interaction terms and inserted in our base model in R to find that only 4 such combinations stood out that may also have some practical relevance to our case-study i.e.,

1. *retire : income*
2. *othdebt : marital*
3. *income : othdebt*
4. *retire : homeown*

We build a model with all these variables and found following results (Fig. 11).

```
Call:
glm(formula = default ~ gender + age + ed + retire + income +
    creddebt + othdebt + marital + homeown + retire:income +
    othdebt:marital + income:othdebt + retire:homeown, family = binomial,
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3754  -0.6954  -0.1904   0.7692   3.6109

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.8452052  0.2880067   6.407 1.49e-10 ***
gender          -0.0494570  0.1004819  -0.492  0.62258
age             -0.0861809  0.0048477 -17.778  < 2e-16 ***
ed               0.0846609  0.0164282   5.153 2.56e-07 ***
retire           1.1842350  0.5321087   2.226  0.02604 *
income          -0.0225300  0.0027777  -8.111 5.02e-16 ***
creddebt         0.5172513  0.0356688  14.502  < 2e-16 ***
othdebt          0.1151492  0.0278138   4.140 3.47e-05 ***
marital          0.1718021  0.1301233   1.320  0.18673
homeown         -0.3093033  0.1067387  -2.898  0.00376 **
retire:income   -0.0409303  0.0199167  -2.055  0.03987 *
othdebt:marital -0.0521210  0.0234971  -2.218  0.02654 *
income:othdebt   0.0002133  0.0001436   1.485  0.13752
retire:homeown  -1.3499949  0.6815927  -1.981  0.04763 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3718.6  on 2720  degrees of freedom
Residual deviance: 2452.7  on 2707  degrees of freedom
AIC: 2480.7

Number of Fisher Scoring iterations: 7
```

Figure 11: Iteration 3 Model Summary

Although the **AIC: 2480.7** and **Residual Deviance: 2452.7**, have increased in comparison to Iteration 2 model, we cannot ignore the possible effects of the interaction terms in the model building step as some relevant combined underlying feature may become significant.

Since now all the assumptions have been considered and the necessary features added to the feature, we can proceed with making our model parsimonious and effective.

iv. Iteration 4 (Rule of Parsimony)

Here we removed the insignificant predictors from Iteration 3 (say *Itr model 3*) model step-by-step in order (*gender>marital>income:othdebt*) each time modelling and depending upon their p values in each step. Following metrics were observed for each model:

1. *Itr model 3 – gender*; Residual Deviance: 2453, AIC: 2479.
2. *Model in 1 – marital;* Residual Deviance*:* 2454.6, AIC: 2478.6.
3. *Model in 2 – income:othdebt;* Residual Deviance: 2457.4, AIC: 2479.4.

```
glm(formula = default ~ age + ed + retire + income + creddebt +
    othdebt + homeown + retire:income + othdebt:marital + retire:homeown,
    family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2210  -0.6959  -0.1917   0.7738   3.9688

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.885917   0.280220   6.730 1.70e-11 ***
age             -0.088142   0.004692 -18.787  < 2e-16 ***
ed               0.081379   0.016314   4.988 6.09e-07 ***
retire           1.330378   0.519911   2.559  0.01050 *
income          -0.019940   0.002245  -8.881  < 2e-16 ***
creddebt         0.502929   0.034287  14.668  < 2e-16 ***
othdebt          0.133836   0.019314   6.929 4.23e-12 ***
homeown         -0.300591   0.106020  -2.835  0.00458 **
retire:income   -0.042247   0.019591  -2.156  0.03105 *
othdebt:marital -0.035403   0.017725  -1.997  0.04578 *
retire:homeown  -1.339768   0.682624  -1.963  0.04968 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3718.6  on 2720  degrees of freedom
Residual deviance: 2457.4  on 2710  degrees of freedom
AIC: 2479.4

Number of Fisher Scoring iterations: 7
```

Figure 12: Final model summary

At this step we also explored multiple models by eliminating barely significant variables, but realized no real improvement in AIC and Deviance metrics. Hence, we declare Iteration 4 Model as our final model with AIC 2479.4 and Deviance 2457.4.

```
> # check residuals and outliers
> sum(rstandard(final_model3)>2.5) # 11 i.e. 0.4 % < 1%
[1] 11
> sum(rstandard(final_model3)>2)# 43 i.e. 1.5 % < 5 %
[1] 43
> # so no residuals problem
> sum(cooks.distance(final_model3)>1)
[1] 0
```

Figure 13: Checking Assumptions validity for final model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1220  331
         1  270  900

               Accuracy : 0.7791
                 95% CI : (0.7631, 0.7946)
    No Information Rate : 0.5476
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5523

 Mcnemar's Test P-Value : 0.01439

            Sensitivity : 0.7311
            Specificity : 0.8188
         Pos Pred Value : 0.7692
         Neg Pred Value : 0.7866
             Prevalence : 0.4524
         Detection Rate : 0.3308
   Detection Prevalence : 0.4300
      Balanced Accuracy : 0.7750

       'Positive' Class : 1
```

Figure 14: Confusion Matrix and Statistics for Final model

Fig. 13 and 14 shows that the final model meets all the requirements for Logistic regression analysis and also provides the classification metrics such as Accuracy (PAC): 77.91 %, Sensitivity: 73.11 % and Specificity: 81.88%. Further evaluation is given in the Result and Evaluation phase.

```
> exp(coef(final_model3))
    (Intercept)             age              ed          retire          income
      6.5923968       0.9156312       1.0847818       3.7824746       0.9802570
       creddebt         othdebt        homeown retire:income othdebt:marital
      1.6535579       1.1432050       0.7403808       0.9586328       0.9652164
retire:homeown
      0.2619064
```

Figure 15: Odds ratios for the final model

Fig.15 and 16 shows the relationships between the Specificity and Sensitivity, and the prediction probabilities for binary classification respectively.

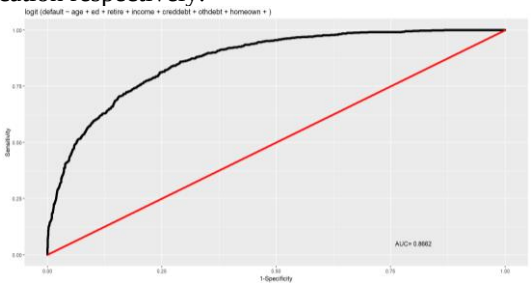


Figure 16: ROC curve for final model

```
> PseudoR2(final_model3, which = "all")
      McFadden     McFaddenAdj        CoxSnell       Nagelkerke   AldrichNelson
     0.3391549       0.3332387       0.3709202       0.4978570       0.3167052
VeallZimmermann          Efron McKelveyZavoina            Tjur             AIC
     0.5484480       0.3967108       0.6659562       0.3948759    2479.4070778
           BIC           logLik          logLik0              G2
  2544.4033800   -1228.7035389   -1859.2914236    1261.1757694
```

Figure 18: PseudoR2 shows the Analogous Goodness of fit for classification

```
> hl <- hoslem.test(mydata$default,fitted(final_model3),g =10 )
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  mydata$default, fitted(final_model3)
X-squared = 8.428, df = 8, p-value = 0.3928
```

Figure 19: Goodness of Fit for Classification

IV. Results and Evaluation

A. *Time Series Analysis Use case*

The final model for SARIMA turned out to be model #4 with (p,d,q)(P,D,Q) values (3,1,4), (1,2,3) and AIC: 5581.62 and RMSE: 2194.99. This model had improved model accuracy and fit and satisfied all assumptions

B. *Logistic Regression Use case*

We observed several models and found that our final logistic regression model presented an Accuracy of 77.91 %, Sensitivity of 73.1% and Specificity 81.8% i.e. the model can correctly classify 77.9% of both the classes while it can classify the "*default: 1" category with* 73.1% accuracy and "*default: 0*" category with 81.8% accuracy.

The evaluation of final model's Odds ratios (Fig.15) informs that for every 1000 units increase (thousands) in *creditdebt* the odds of getting a *default* :1 increases by 1.6 times while every unit increase (years) in *othdebt* increases the odds of getting a *default: 1.*

Finally, the ROC plots and Jitter-plots in Fig. 16 and 17 provides us with the relationship between Sensitivity and Specificity, and prediction probabilities for classification of binary classes respectively.