Terminal Assignment Based Assessment-Statistics

John Maruthukunnel Jacob
Department of Computing
National College of Ireland
Dublin, Ireland
x21138494@student.ncirl.ie

***Abstract*—In the first part of this assignment the aim is to comprehend how the number of private car registrations in Ireland has changed in the time period 1995-2022. The intent is to visualize and analyze the trends and find the best predictive model for six periods forward using time series analysis.**
In the second part, logistic regression is used to build a predictive model to best predict the categorical dependent variable (if a customer has loan default or not) in the dataset Default.

***Index Terms*—Time series, Logistic Regression, Modelling, Correlation**

I. Time Series Analysis

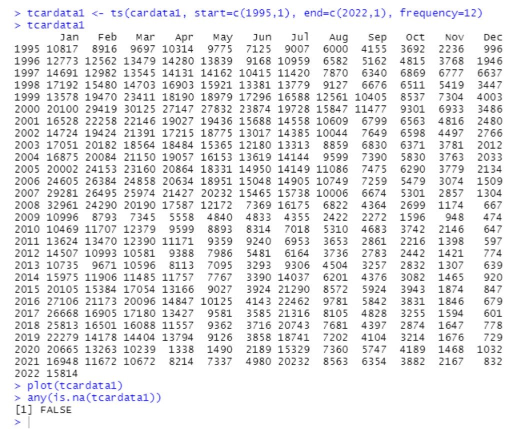
A. Introduction

A time series is a grouping of observations on a variable estimated at progressive moments or over progressive time frames. The observations might be taken each hour, day, week, month, or year or at some other standard span(cyclical,seasonal etc.). Mathematically, it is the noted observations y_1, y_2, y_3, ..., y_n at time frames t_1, t_2, t_3, ..., t_n.
For the given dataset, quantitative forecasting using numerical data regarding historical data is given, and the observed pattern will continue.

B. Preliminary Assessment

The CarRegistrations.csv file downloaded from Moodle contains data about private car registrations in Ireland over the period January 1995 to January 2022. Initially the dataset was loaded into R and stored in a timeseries object using ts() fucntion which is shown in fig. 1.

As the data is monthly observations, we used the frequency=12. Column headers for the dataset was added (TimePeriod, NoRegistrations). The dataset was checked for any null values using any(is.any()) function and none was found. The basic plot for the time series is plotted and shown in fig.2



```
> tcardata1 <- ts(cardata1, start=c(1995,1), end=c(2022,1), frequency=12)
> tcardata1
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
1995 10817  8916  9697 10314  9775  7125  9007  6000  4155  3692  2236   996
1996 12773 12562 13479 14280 13839  9168 10959  6582  5162  4815  3768  1946
1997 14691 12982 13545 14131 14162 10415 11420  7870  6340  6869  6777  6637
1998 17192 15480 14703 16903 15921 13381 13779  9127  6676  6511  5419  3447
1999 13578 19470 23411 18190 18979 17296 16588 12561 10405  8537  7304  4003
2000 20100 29419 30125 27147 27832 23874 19728 15847 11477  9301  6933  3486
2001 16528 22258 22146 19027 19436 15688 14558 10609  6799  6563  4816  2480
2002 14724 19424 21391 17215 18775 13017 14385 10044  7649  6598  4497  2766
2003 17051 20182 18564 18484 15365 12180 13313  8859  6830  6371  3781  2012
2004 16875 20084 21150 19057 16153 13619 14144  9599  7390  5830  3763  2033
2005 20002 24153 23160 20864 18331 14950 14149 11086  7475  6290  3779  2134
2006 24605 26384 24858 20634 18951 15048 14905 10749  7259  5479  3074  1509
2007 29281 26495 25974 21427 20232 15465 15738 10006  6674  5301  2857  1304
2008 32961 24290 20190 17587 12172  7369 16175  6822  4364  2699  1174   667
2009 10996  8793  7345  5558  4840  4833  4355  2422  2272  1596   948   474
2010 10469 11707 12379  9599  8893  8314  7018  5310  4683  3742  2146   647
2011 13624 13470 12390 11171  9359  9240  6953  3653  2861  2216  1398   597
2012 14507 10993 10581  9388  7986  5481  6164  3736  2783  2442  1421   774
2013 10735  9671 10596  8113  7095  3293  9306  4504  3257  2832  1307   639
2014 15975 11906 11485 11757  7767  3390 14037  6201  4376  3082  1465   920
2015 20105 15384 17054 13166  9027  3924 21290  8572  5924  3943  1874   847
2016 27106 21173 20096 14847 10125  4143 22462  9781  5842  3831  1846   679
2017 26668 16905 17180 13427  9581  3585 21316  8105  4828  3255  1594   601
2018 25813 16501 16088 11557  9362  3716 20743  7681  4397  2874  1647   778
2019 22279 14178 14404 13794  9126  3858 18741  7202  4104  3214  1676   729
2020 20665 13263 10239  1338  1490  2189 15329  7360  5747  4189  1468  1032
2021 16948 11672 10672  8214  7337  4980 20232  8563  6354  3882  2167   832
2022 15814
> plot(tcardata1)
> any(is.na(tcardata1))
[1] FALSE
>
```

Fig. 1. Data Set Summary.

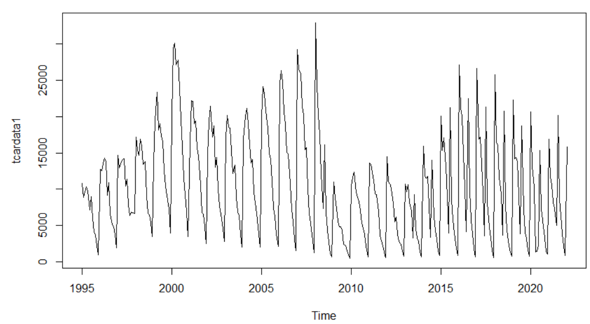


Fig. 2. Graph plot of Ireland's Private Car Registration.

C. Components of Time Series

Time Series Analysis uses various visualization techniques to represent the dataset more clearly as shown below:

Level

From fig. 2, we observe that there is a level which is not increasing for the number of private car registrations in Ireland.

Trend

From fig. 2, we observe that there is an increasing trend from 1995-2000, 2002-2008, 2013-2017 and a decreasing trend from 2000-2002, 2008-2013, 2017-2022. Overall there has been a sharp trend throughout the years.

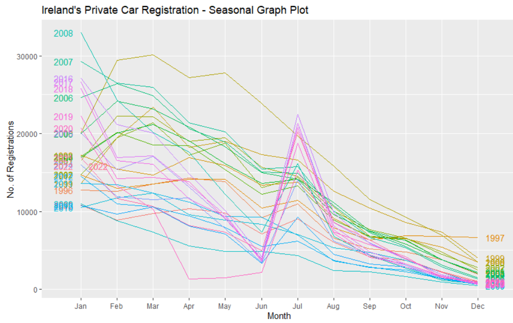


Fig. 3. Ireland's Private Car Registration - Seasonal Graph Plot.

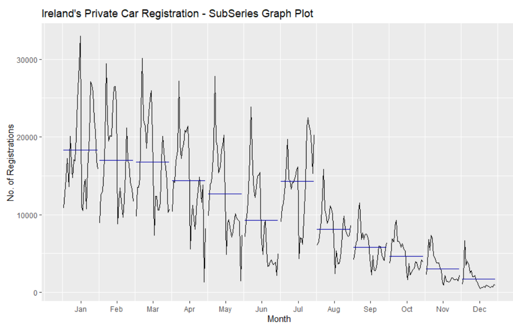


Fig. 4. Ireland's Private Car Registration - Seasonal Sub-series Graph Plot.

Seasonal

From fig.3 and fig. 4, we can observe that the data has a strong seasonal pattern. We observe that most registrations occur in January while the month December sees the least. There is a sudden increase in registrations over the period June and decrease in August. It was also observed that the data is not cyclic.

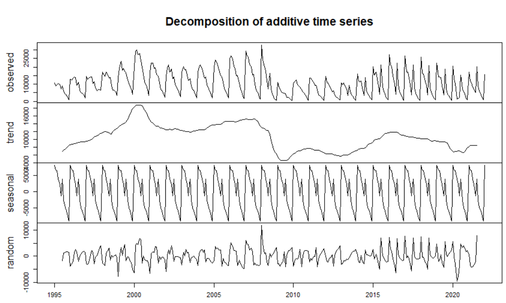
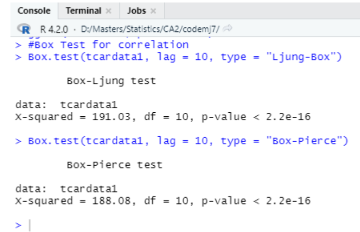


Fig. 5. Ireland's Private Car Registration - Decomposition Plot.

The dataset shows a strong seasonal pattern compared to trend pattern and level as shown in fig. 5




```
> #Box Test for correlation
> Box.test(tcardata1, lag = 10, type = "Ljung-Box")

        Box-Ljung test

data:  tcardata1
X-squared = 191.03, df = 10, p-value < 2.2e-16

> Box.test(tcardata1, lag = 10, type = "Box-Pierce")

        Box-Pierce test

data:  tcardata1
X-squared = 188.08, df = 10, p-value < 2.2e-16

>
```

Fig. 6. Box-Ljung test and Box-Pierce test

Auto-correlation and White Noise

Box-Ljung test and box-Pierce test was done to check for white noise and correlation. The observed values as shown in Fig 6 are X-Squared value of 191.03 and p-value $<$ 2.2e-16(Box-Ljung test) and X-Sqaured value of 188.08 and p-value $<$ 2.2e-16(Box-Pierce test).
So we have rejected the null hypothesis that no white noise is present and accepted H1 – there is correlation.

D. Categorical Time Series Models

Exponential Smoothing

There are three main models in exponential smoothing which are simple exponential model (no trend or seasonal component), Holt exponential smoothing(with level and trend) and Holt-Winters exponential smoothing (level ,trend, seasonal components are present). Since our dataset contains components of level, trend and seasonal, we have selected Holt-Winters exponential smoothing model (Additive, Multiplicative and Damped Multiplicative) as best suited.

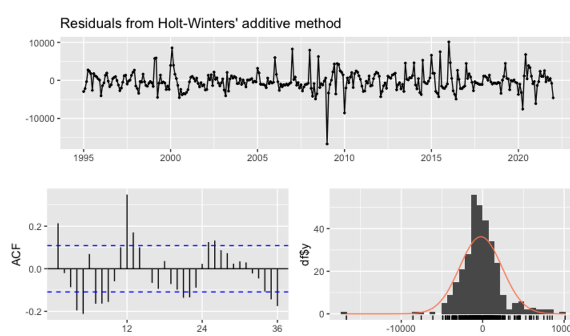


Fig. 7. Residual Diagnostics of Additive model

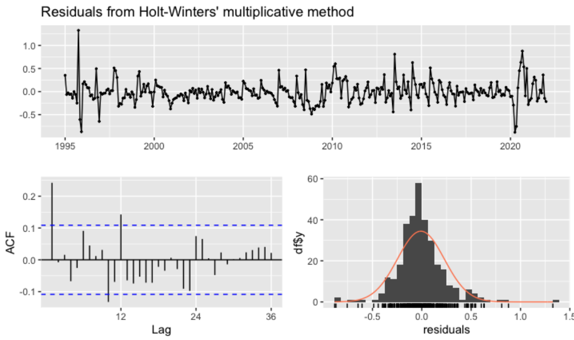


Fig. 8. Residual Diagnostics of Multiplicative model

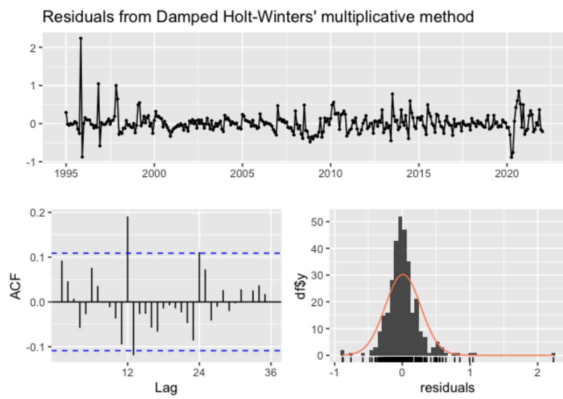


Fig. 9. Residual Diagnostics of Damped Multiplicative model

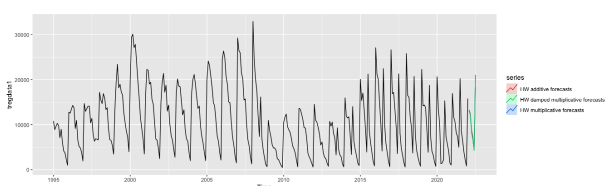


Fig. 10. Prediction Graph for Next Six Period

The residual plots for Holt Winter's Additive model(fig 7), Multiplicative model(fig 8) and Damped Multiplicative model(fig 9) is shown above. The Ljung-Box test results are shown below in fig.

```
        Ljung-Box test

data:  Residuals from Holt-Winters' additive method
Q* = 158.1, df = 8, p-value < 2.2e-16

        Ljung-Box test

data:  Residuals from Holt-Winters' multiplicative method
Q* = 56.242, df = 8, p-value = 2.53e-09

        Ljung-Box test

data:  Residuals from Damped Holt-Winters' multiplicative method
Q* = 39.267, df = 7, p-value = 1.738e-06
```

Fig. 11. Ljung-Box test

From the above model we can observe that additive model has alpha = 0.4944, beta = 0.0028, gamma = 0.465 and RMSE = 2604.6.
We can also observe the values of Multiplicative model having alpha = 0.2556, beta = 0.003, gamma = 0.4619 and RMSE = 2291.275 while Damped Mulitiplicative model has alpha = 0.2532, beta = 1e-04, gamma = 0.4885 and RMSE = 2277.514.

The model which has the least value for beta and RMSE is chosen as the better fit. From the above observed data we can see that the Holt Winter's Multiplicative model has the lesser beta and RMSE values of all models. Holt Winter's Multiplicative Model is a better fit because it shows the seasonality of the data with no big slope changes.

ARIMA AND SARIMA MODELS

ARIMA

ARIMA and SARIMA are models implemented to better fit stationary time-series datasets. In the previous section we observed that the p value is small and the dataset does not contain white noise using the ACF and PACF plot as shown in Fig.12

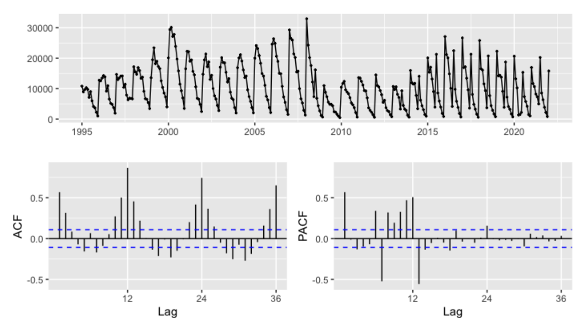


Fig. 12. ACF and PACF Plots

The functions *acf*, *pacf*, and *ggtsdisplay* were used in R to plot the ACF and PACF plots.

The auto.arima function was used in R studio to get the best model fit for the time series. The figure Fig. 13 shows the residual plot for the best fit model from ARIMA model and the best fit model found was (1,1,1)(1,1,2)[12]

The Fig.14 shows the accuracy summary of the auto fit ARIMA model (1,1,1)(1,1,2)[12] with RMSE = 2282. 768, p-value = 0.002692 and ACF1 = -0.07412356

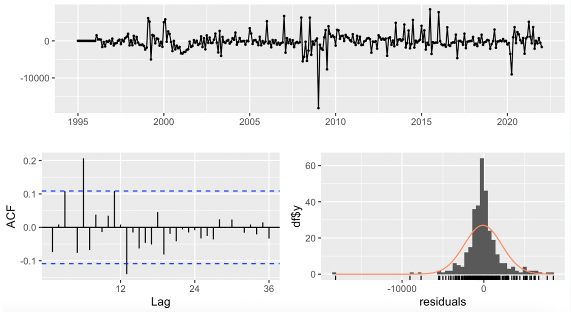


Fig. 13. ARIMA Residual Plot

```
Coefficients:
         ar1      ma1    sar1     sma1
      0.6716  -0.9553  0.6639  -0.8824
s.e.  0.0915   0.0569  0.0928   0.0636

sigma^2 = 5498650:  log likelihood = -2864.49
AIC=5738.98   AICc=5739.18   BIC=5757.7

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -109.3054 2282.768 1382.935 -2.149911 20.40798 0.6330942 -0.07412356
```

Fig. 14. ARIMA Model Accuracy Summary

The ARIMA model (1,1,1)(1,1,2)[12] was found to be a better fit and is used to build the SARIMA model in the section below.

SARIMA

Seasonal ARIMA/ SARIMA, is an extensive modelling of ARIMA that is better suited for univariate time series data with a seasonal component. The SARIMA model was created for the model chosen from the above section (1,1,1)(1,1,2)[12].

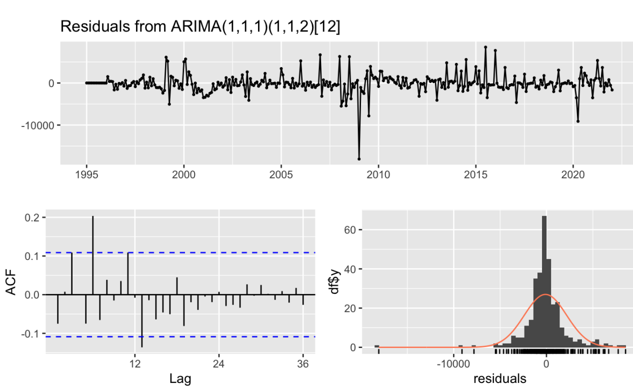


Fig. 15. Seasonal ARIMA Model Residual plot

```
> summary(fit_sarima1t)
Series: tregdata1
ARIMA(1,1,1)(1,1,2)[12]

Coefficients:
         ar1      ma1    sar1     sma1     sma2
      0.6741  -0.9576  0.6156  -0.8166  -0.0428
s.e.  0.0813   0.0489  0.1337   0.1416   0.0810

sigma^2 = 5512365:  log likelihood = -2864.35
AIC=5740.71   AICc=5740.98   BIC=5763.17

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -111.431 2281.899 1379.051 -2.200514 20.33693 0.6313162 -0.07501692
> accuracy(fit_sarima1t)
                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -111.431 2281.899 1379.051 -2.200514 20.33693 0.6313162 -0.07501692
> checkresiduals(fit_sarima1t)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)(1,1,2)[12]
Q* = 41.08, df = 19, p-value = 0.002355

Model df: 5.   Total lags used: 24

>
```

Fig. 16. Seasonal ARIMA Model Accuracy Summary

THe SARIMA model residual plots were plotted and shown in Fig.15. From the Fig.16 we can see the accuracy of the SARIMA model, with p-value = 0.002355, RMSE = 2281.899 and ACF1 = -0.075016.
So we can conclude from the above accuracy measures that the SARIMA model is a better fit for the Time Series.

Simple Time Series

The Simple Time Series is the model used for straightforward time series modelling and it can show less accuracy based on trend and seasonality. However Seasonal Naive and Drift models can be used to get better accuracy for time series with seasonality or trend. In our dataset we have strong seasonality and sharp trend(refer Fig 3 and Fig4). So we have done the residual diagnostics of Seasonal Naive and Drift model to compare with the model accuracies obtained in the previous sections.

The below figures Fig.17 and Fig.18 shows the accuracy summary and residual plots for Seasonal Naive model.

```
> fcast.snaive <- snaive(tcardata1, h=6)
> summary(fcast.snaive)

Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = tcardata1, 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
> plot(fcast.snaive)
> checkresiduals(fcast.snaive)

        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 554.57, df = 24, p-value < 2.2e-16

Model df: 0.   Total lags used: 24
```

Fig. 17. Seasonal Naive Model Accuracy Summary

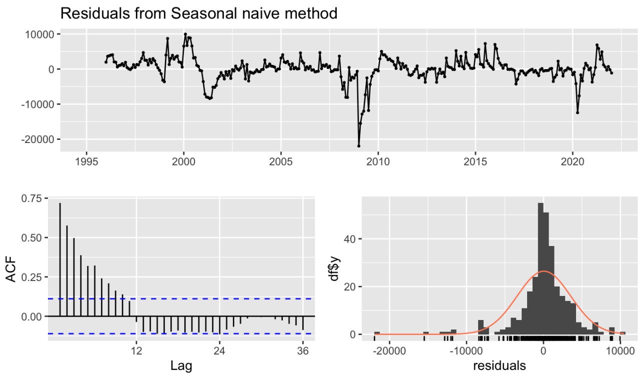
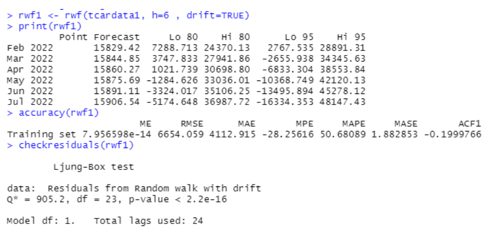


Fig. 18. Seasonal Naive Model Residual Plot

The accuracy summaries and residual plots for Drift model is shown in Fig.19 and Fig.20 .



```
> rwf1 <- rwf(tcardata1, h=6 , drift=TRUE)
> print(rwf1)
         Point Forecast     Lo 80    Hi 80      Lo 95    Hi 95
Feb 2022       15829.42  7288.713 24370.13   2767.535 28891.31
Mar 2022       15844.85  3747.833 27941.86  -2655.938 34345.63
Apr 2022       15860.27  1021.739 30698.80  -6833.304 38553.84
May 2022       15875.69 -1284.626 33036.01 -10368.749 42120.13
Jun 2022       15891.11 -3324.017 35106.25 -13495.894 45278.12
Jul 2022       15906.54 -5174.648 36987.72 -16334.353 48147.43
> accuracy(rwf1)
                       ME     RMSE      MAE       MPE     MAPE     MASE       ACF1
Training set 7.956598e-14 6654.059 4112.915 -28.25616 50.68089 1.882853 -0.1999766
> checkresiduals(rwf1)

        Ljung-Box test

data:  Residuals from Random walk with drift
Q* = 905.2, df = 23, p-value < 2.2e-16

Model df: 1.   Total lags used: 24
```

Fig. 19. Drift Model Accuracy Summary

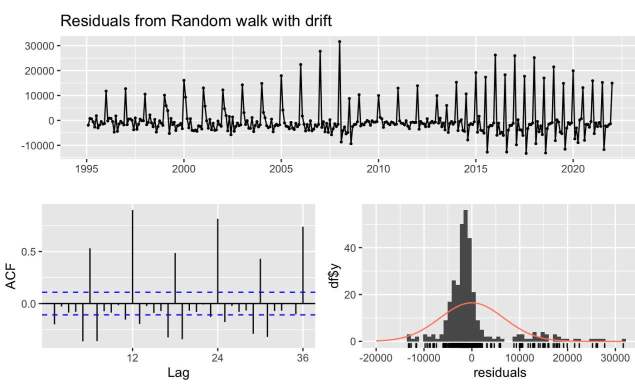
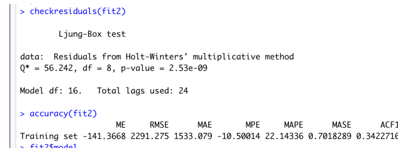


Fig. 20. Drift Model Residual Plot

From comparing the p-values, RMSE, ACF1 values of both models with each other and the previous models we observe that they are not as better suited model as SARIMA model.

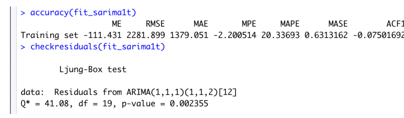
E. CONCLUSION

From the above models we have build in this assignment we can select an optimum model to forecast for the next 6 months for the given dataset. The accuracy tests done in R for the models is shown below in the Fig.21, Fig. 21 and Fig. 22



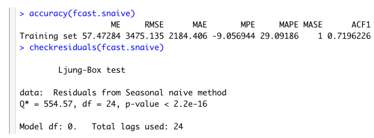
```
> checkresiduals(fit2)

        Ljung-Box test

data:  Residuals from Holt-Winters' multiplicative method
Q* = 56.242, df = 8, p-value = 2.53e-09

Model df: 16.   Total lags used: 24

> accuracy(fit2)
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -141.3668 2291.275 1533.079 -10.50014 22.14336 0.7018289 0.3422716
> fit2$model
```

Fig. 21. Accuracy test for HW Multiplicative Model



```
> accuracy(fit_sarima1t)
                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -111.431 2281.899 1379.051 -2.200514 20.33693 0.6313162 -0.07501692
> checkresiduals(fit_sarima1t)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)(1,1,2)[12]
Q* = 41.08, df = 19, p-value = 0.002355
```

Fig. 22. Accuracy test for SARIMA Model



```
> accuracy(fcast.snaive)
                   ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 57.47284 3475.135 2184.406 -9.056944 29.09186    1 0.7196226
> checkresiduals(fcast.snaive)

        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 554.57, df = 24, p-value < 2.2e-16

Model df: 0.   Total lags used: 24
```

Fig. 23. Accuracy test for Seasonal Naive Model

From the above accuracy measures, the best RMSE value is 2281.899 and MAPE value is 20.3369 we can conclude that the SARIMA model outperformed the other models and is better suited to predict the car registrations for the next 6 months. The SARIMA Model 6 month forecast for the dataset is given in Fig.24

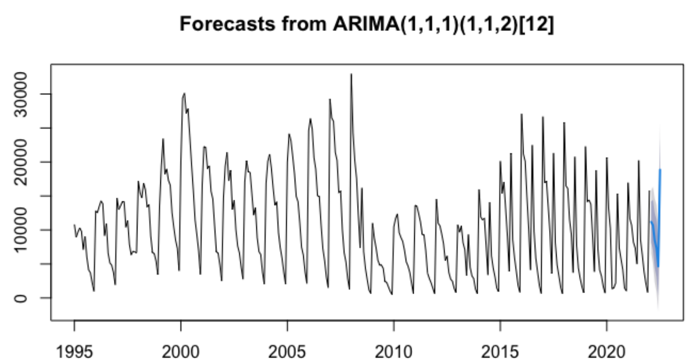


Fig. 24. 6 Month Forecast for SARIMA Model

II. LOGISTIC REGRESSION

A. INTRODUCTION

Logistic Regression is an analytics approach to determine the effects of a collection of independent variables on a dependent variable which is finite or categorical: either X or Y (binary regression) or a range of finite options A, B, C or D (multinomial regression). It aids in predictive analytics and modeling categorical variables with binary, ordinal and nominal values. In this assignment we are creating a logistic regression model to predict whether a customer has defaulted (dependent variables) based on independent variables given in fig.25

B. DATA DESCRIPTION

The given dataset default.csv contains records of 2721 customers. The dataset contains 10 variables of which 4 are continuous and 6 are categorical. The dependent variable is default as shown in fig.25

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	1	0		None	None	8	Right	Nominal	Input
2	age	Numeric	2	0		None	None	8	Right	Scale	Input
3	ed	Numeric	2	0		None	None	8	Right	Nominal	Input
4	retire	Numeric	1	0		None	None	8	Right	Nominal	Input
5	income	Numeric	4	0		None	None	8	Right	Scale	Input
6	creddebt	Numeric	10	6		None	None	12	Right	Scale	Input
7	othdebt	Numeric	10	6		None	None	12	Right	Scale	Input
8	default	Numeric	1	0		None	None	8	Right	Nominal	Input
9	marital	Numeric	1	0		None	None	8	Right	Nominal	Input
10	homeown	Numeric	1	0		None	None	8	Right	Nominal	Input

Fig. 25. Variable Table

SPSS and RStudio were used to calculate descriptive statistics for all variables in the dataset. From the fig. 25 we can see the skewness is high for income, creddebt and othdebt. The positive skewness statistic for these variables suggest the presence of positive outliers for the variables.The boxplots of these three variables are shown in Fig.27

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
gender	2721	0	1	.52	.500	.250	-.070	.047	-1.997	.094
age	2721	18	79	43.91	17.795	316.660	.308	.047	-1.096	.094
ed	2721	6	23	14.76	3.271	10.699	-.050	.047	-.623	.094
retire	2721	0	1	.11	.317	.101	2.437	.047	3.943	.094
income	2721	9	1073	54.69	60.138	3616.530	6.046	.047	69.742	.094
creddebt	2721	.001364	109.072596	2.20815118	4.33452523	18.788	9.675	.047	172.126	.094
othdebt	2721	.016704	141.459150	3.92953093	6.02625176	36.316	8.236	.047	136.206	.094
default	2721	0	1	.43	.495	.245	.283	.047	-1.921	.094
marital	2721	0	1	.47	.499	.249	.101	.047	-1.991	.094
homeown	2721	0	1	.63	.482	.232	-.558	.047	-1.690	.094
Valid N (listwise)	2721									

Fig. 26. Descriptive Statistics

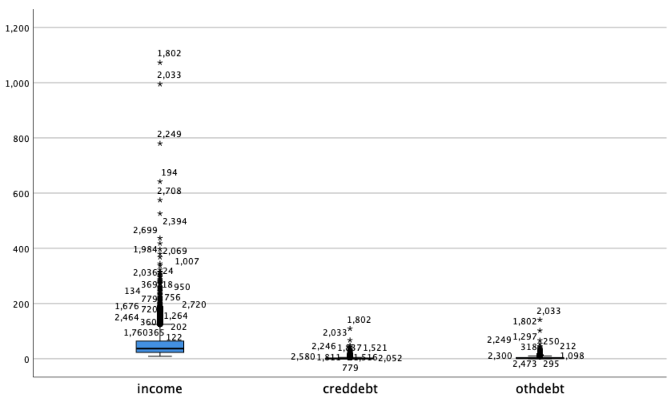


Fig. 27. Boxplot- Income, Creddebt, Othdebt

The Pearson correlation was done in SPSS and shown in fig.28 . From the correlation table we can see that the variables income, creddetb and othdebt is most correlated with the dependent variable default.

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. 28. Pearson Correlation table

C. LOGISTIC REGRESSION ASSUMPTIONS

- The outcome variable of the dataset should be binary. The outcome variable default in the given dataset is binary.
- Then dataset should contain appropriate sample size and the given dataset has 2721 records.
- The continous variables should be linear and normally distributed. To meet this requirement the log of the continous variables were taken.
- Absence of outliers-the outliers of income variable was removed from the dataset.

D. MODEL CREATION

MODEL 1

In the first model we have taken all variables for regression.
Model 1 : default gender+age+ed+retire+income+creddebt+othdebt +marital+homeown
The summary of the model 1 is shown in Fig.29

```
> summary(modelone)

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

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 ***
gender1     -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 ***
retire1     -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 ***
marital1    -0.019930   0.100480  -0.198 0.842774
homeown1    -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
```

Fig. 29. Model 1 Summary

From the summary we can observe the variables age, ed, income, creddebt, othdebt and homeown based on the p-values. The variables gender, retire and marital are insignificant as their respective p-values>0.05.

MODEL 2

In the second model we have removed the insignificant variables gender, retire and marital.
Model 2 : default age+ed+income+creddebt+othdebt+homeown
The summary of the model 2 is shown in Fig.30

```
> summary(modeltwo)

Call:
glm(formula = default ~ age + ed + income + creddebt + othdebt +
    homeown, family = "binomial", data = d_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0168  -0.7007  -0.2167   0.7743   4.0556

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.907155   0.274496   6.948 3.71e-12 ***
age         -0.087702   0.004055 -21.630  < 2e-16 ***
ed           0.081881   0.016267   5.033 4.82e-07 ***
income      -0.019619   0.002190  -8.957  < 2e-16 ***
creddebt     0.492357   0.033525  14.686  < 2e-16 ***
othdebt      0.114913   0.016904   6.798 1.06e-11 ***
homeown1    -0.357791   0.103640  -3.452 0.000556 ***
---
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: 2472.1  on 2714  degrees of freedom
AIC: 2486.1

Number of Fisher Scoring iterations: 6

> |
```

Fig. 30. Model 2 Summary

MODEL 3

In the Final model we have taken the variables:
Model 3 : default age+ed+log_income+log_creddebt+ log_othdebt+homeown
The summary of the final model is shown in Fig.31

```
> summary(model3)

Call:
glm(formula = default ~ age + ed + log_income + log_creddebt +
    log_othdebt + homeown, family = "binomial", data = lr_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3707  -0.7479  -0.2333   0.7816   3.1741

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.319014   0.411196  12.935  < 2e-16 ***
age          -0.084326   0.003956 -21.317  < 2e-16 ***
ed            0.067003   0.015936   4.204 2.62e-05 ***
log_income   -0.886548   0.110712  -8.008 1.17e-15 ***
log_creddebt  0.813576   0.060625  13.420  < 2e-16 ***
log_othdebt   0.442990   0.069401   6.383 1.74e-10 ***
homeown1     -0.343263   0.102292  -3.356 0.000792 ***
---
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: 2584.2  on 2714  degrees of freedom
AIC: 2598.2

Number of Fisher Scoring iterations: 5
```

Fig. 31. Final Model Summary

```
> confusionMatrix(data = pred, lr_data$default)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1223  266
         1  328  904

               Accuracy : 0.7817
                 95% CI : (0.7657, 0.7971)
    No Information Rate : 0.57
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5575

 Mcnemar's Test P-Value : 0.01232

            Sensitivity : 0.7885
            Specificity : 0.7726
         Pos Pred Value : 0.8214
         Neg Pred Value : 0.7338
             Prevalence : 0.5700
         Detection Rate : 0.4495
   Detection Prevalence : 0.5472
      Balanced Accuracy : 0.7806

       'Positive' Class : 0
```

Fig. 32. Confusion Matrix for Final Model

```
> coef(model3)
 (Intercept)          age           ed   log_income log_creddebt  log_othdebt
  5.90424987  -0.08290779   0.06653798  -1.07873936   0.69039381   0.44448138
    homeown1
 -0.34617445
> exp(coef(model3))
 (Intercept)          age           ed   log_income log_creddebt  log_othdebt
 366.5921308    0.9204360    1.0688016    0.3400239    1.9945008    1.5596811
    homeown1
   0.7073891
```

Fig. 33. Coefficients and Odds Ratio- Final Model

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1300.908	9	.000
	Block	1300.908	9	.000
	Model	1300.908	9	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2417.675[a]	.380	.510

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	7.873	8	.446

Fig. 34. Tests for Final Model

The logistic regression assumptions were tested for the final model and the correlation between the continous variables log_income, log_creddebt, log_othdebt is less than 0.70 and we can observe that value of VIF < 5 ,which confirms the absence of multicollinearity in predictor variables.The tests are shown in fig. 35

```
> cor.test(clean_data$log_othdebt , clean_data$log_creddebt, method = "pearson")

        Pearson's product-moment correlation

data:  clean_data$log_othdebt and clean_data$log_creddebt
t = 40.271, df = 2406, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6100577 0.6578111
sample estimates:
      cor
0.6345396

> cor.test(clean_data$log_income , clean_data$log_creddebt, method = "pearson")

        Pearson's product-moment correlation

data:  clean_data$log_income and clean_data$log_creddebt
t = 28.948, df = 2406, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4780108 0.5372877
sample estimates:
     cor
0.508251

> cor.test(clean_data$log_othdebt , clean_data$log_income, method = "pearson")

        Pearson's product-moment correlation

data:  clean_data$log_othdebt and clean_data$log_income
t = 33.931, df = 2406, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5412588 0.5953199
sample estimates:
      cor
0.5689036

   vif(model3)
        age          ed   log_income log_creddebt  log_othdebt     homeown
   1.213830    1.077975     1.941607     1.861699     1.934981    1.013361
```

Fig. 35. Assumption tests for Final Model

E. CONCLUSION

From the above summary we can observe that Cox and Snell R Square value and Nagelkerke R square values are greater than 0.3 and Odds ratio of *log_creddebt* and *log_othdebt* are 1.994 and 1.56 respectively which is good for fit model.
The overall accuracy of final model = 78.17% with sensitivity = 78.85% and specificity = 77.26%.