Time Series Analysis and Logistic Regression

1st Rutuja Dinesh Mehta
MSc in Data Analytics
National College Of Ireland
Dublin, Ireland
x20129751@student.ncirl.ie

***Abstract*—Time series analysis is used to analyze, visualize and gain statistics from the time series data which has the data collected at different intervals of time. These are usually the points collected at the successive measurements from the same source of data, used for tracking the change along with time. Logistic Regression, an extension for Linear Regression, is used for predicting the probabilities for the classification problems for the dichotomous dependent variable, i.e variable with two possible outcomes. This paper demonstrates the time series analysis for predicting the e-commerce retail sales of the States which are commenced from Q4, 1999. Binary logistic regression is used for predicting the house category, expensive or budget, based on various parameters.**

I. Part A : Time Series Analysis

A. Data Description and Data Understanding

The dataset consists of the e-commerce retail sales of the United States from Quarter 4, 1999 to Quarter 2, 2021. It has total of 87 observations. The dataset is an quarterly based time series, used for predicting the US retail sales of the coming next three quarters measured in billion dollars.

The initial step is plotting a time series graph for understanding the patterns in the data. This then helps in categorizing the data into the observed trend or some seasonal. A forecasting model is then chosen depending upon the observed pattern.

```
> tusa_retail <- ts(usa_retail, start = (c(1999,4)), frequency = 4)
> is.ts(tusa_retail)
[1] TRUE
> plot(tusa_retail, main = "United States E-Commerce Retail Sales")
```

Fig. 1. Time Series Plot

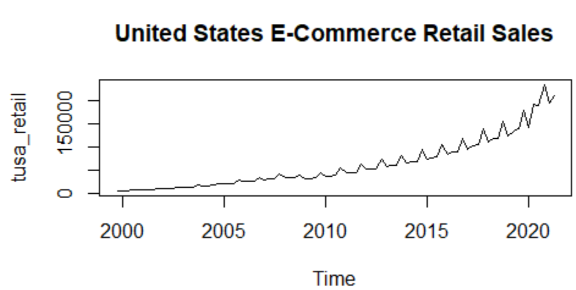

Fig. 2. Time Series Plot

The above graph depicts that using the ts function, both the patterns are observed, trends and the seasonal, it is the combination of both. It can also be concluded that the US retail sales are low in the first two quarters of every year and the sales increase in the third and the fourth quarter.

B. Seasonality

Seasonal plot which is calculated using a seasonal plot shows us the observed pattern for each season, on a quarterly basis here. Seasonal sub-series plot, used to calculate mean for every quarter, denoted with a horizontal line here in the graph. From the graphs, seasonal and the seasonal sub-series, it can be observed that the sales were low in the first two quarters of each year and then increased for the next two quarters, Q3 and Q4. This pattern is observed for the year ranging from 1991 to 2021. Graph below gives us the insights of the observed pattern in the dataset.

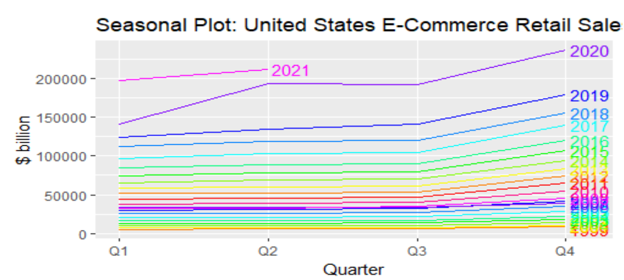

Fig. 3. Seasonal Plot

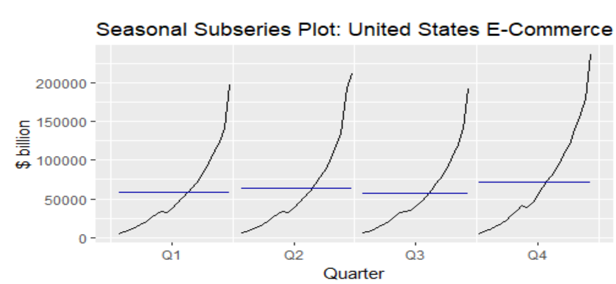

Fig. 4. Seasonal Subseries Plot

C. Seasonal Decomposition

A series is decomposed into a random component, observed component, trend component and a seasonal component by the method of seasonal decomposition. Seasonal decomposition can be either be multiplicative or additive. Additive is observed here. It is a summation of components; trend, seasonal and irregular.

Additive Seasonal Decomposition :

$$\mathbf{Y}_t = \mathbf{Trend}_t + \mathbf{Seasonal}_t + \mathbf{Irregular}_t \tag{1}$$

where, the observations taken at time t are the sum of the factors that contributed to the trend, seasonal and irregular effects.

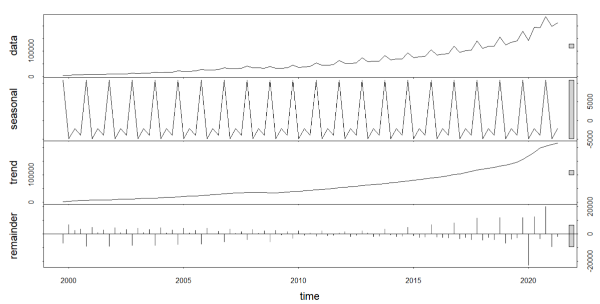

Fig. 5. Seasonal Decomposition

The additive seasonal decomposition model is used when the seasonal variations are independent of the changing time. Here, in the given time series dataset, the variations of the seasonsin the initial period is same as the seasonal variations of the post period, So, it can be said that the additive model for seasonal decomposition fits here.

D. Model Building

a) Simple Exponential Smoothing

Simple exponential smoothing, also known as a single exponential model, is used for forecasting a time series model for the data which doesn't have a trend or seasonality, a data that is univariate. It just requires a single smoothing factor or a smoothing coefficient known as alpha(a).The weighted averages is used for calculating the forecast values.

It can be seen that the RSME value stands out to be 13193.76 and AIC value as 2045.359, which are exceptionally high. The predicted value 208080.5 is same for the next three quarters and isn't following any trend as compared to the previous trend. Hence, the Simple Exponential Smoothing model is dropped.

```
> summary(fc_sex)

Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = simple_ex, h = 3)

  Smoothing parameters:
    alpha = 0.5447

  Initial states:
    l = 6946.7401

  sigma:  13348.08

     AIC     AICc      BIC
2045.359 2045.648 2052.757

Error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 4244.196 13193.76 7275.815 5.423953 10.33686 0.7328429 -0.284854

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       208080.5 190974.3 225186.8 181918.8 234242.3
2021 Q4       208080.5 188601.1 227560.0 178289.3 237871.8
2022 Q1       208080.5 186487.1 229674.0 175056.3 241104.8
```

Fig. 6. Simple Exponential Smoothing Summary

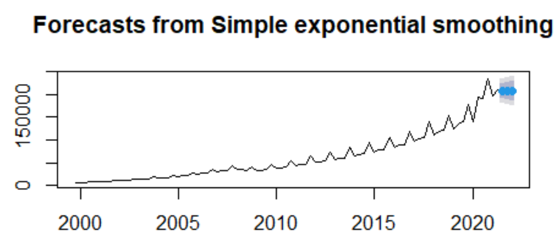

Fig. 7. Simple Exponential Smoothing Plot

b) Seasonal Naive

The seasonal naive method is used to predict the last recorded value from the same season of the year with high seasonality.

```
> seasonalnaive_usaretail <- snaive(tusa_retail, h = 3)
> summary(seasonalnaive_usaretail)

Forecast method: Seasonal naive method

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

Residual sd: 15143.9963

Error measures:
                   ME  RMSE      MAE      MPE     MAPE MASE      ACF1
Training set 9786.711 15144 9928.205 15.44986 15.85113    1 0.8350162

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3         191573 172165.2 210980.8 161891.3 221254.7
2021 Q4         235957 216549.2 255364.8 206275.3 265638.7
2022 Q1         196808 177400.2 216215.8 167126.3 226489.7
```

Fig. 8. Summary for Seasonal Naive Model

The summary gives us the RSME value as 15144, found to be higher than the RSME value of simple exponential smoothing. But it's p-value is low than p<2.2e-16 as per the LJung Box test. The graph depicted also follows a trend with

```
> Box.test(tusa_retail, lag = 1, type = "Ljung")

        Box-Ljung test

data:  tusa_retail
X-squared = 74.561, df = 1, p-value < 2.2e-16
```

Fig. 9. LJung-Box Test for Seasonal Naive Model

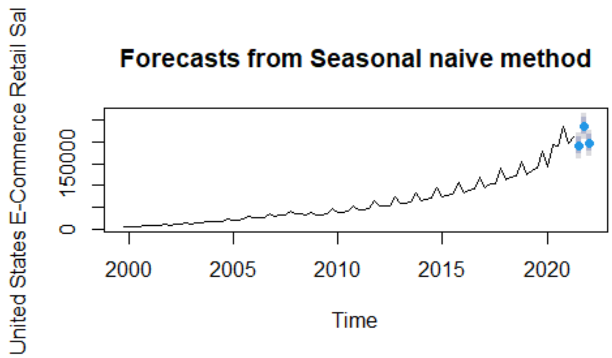

Fig. 10. Forecast for Seasonal Naive Model

the last recorded value. But still this model is dropped and an additive HoltWinters Model will be used for better prediction of RSME and p-values.

c) Additive HoltWinters

Additive HoltWinters is an extension to Holt's Exponential smoothing which is used in capturing seasonality. It is similar to the multiplicative model, the only difference is that the seasonality is considered as additive. It is best fit for data that which has unchangeable trend and seasonality with changing time. The forecasted value is the summation of seasonality components, trend and baseline for each data element. It has the forecast in the curved shape with the seasonal changes.

```
Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = tusa_retail)

Smoothing parameters:
 alpha: 0.6299163
 beta : 0.1357028
 gamma: 1

Coefficients:
          [,1]
a   205360.8400
b     6426.8962
s1     180.4556
s2   31225.7917
s3  -12252.5163
s4    6343.1600

Error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 830.0594 5629.452 2575.731 0.7719355 3.632903 0.2594357 -0.035025

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       211968.2 204789.2 219147.1 200988.9 222947.5
2021 Q4       249440.4 240613.5 258267.3 235940.9 262940.0
2022 Q1       212389.0 201854.8 222923.3 196278.3 228499.8
```

Fig. 11. Summary for Additive HoltWinters Model

The RSME value comes out to be 5629, which can be said as a low and good score. 0.01 is the p-value, less than the

```
> Box.test(HW1_for, lag = 1, type = "Ljung")

        Box-Ljung test

data:  HW1_for
X-squared = 5.9248, df = 1, p-value = 0.01493
```

Fig. 12. LJung-Box Test for Additive HoltWinters Model

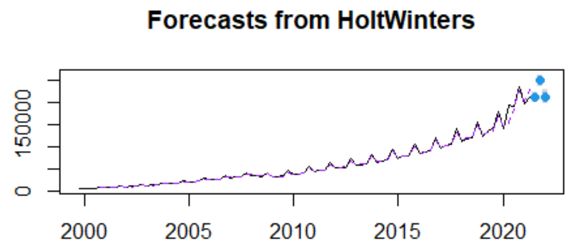

Fig. 13. Forecast for Additive HoltWinters Model

0.05, has a good score as well. This model can be said as a model which follows a good trend and a good pattern.

d) Seasonal ARIMA Model

The most commonly used time series for forecasting a model are ARIMA (Autoregressive integrated moving average) and the seasonal ARIMA. SARIMA is different from the ARIMA as it follows a concept of seasonal trends. A statistical approach is used for time series prediction which has an irregular component with non-zero autocorrections. A fit stattionary model can be built from these models. Since, the dataset of US retail sales includes seasonal and the non-seasonal components, seasonal ARIMA model is used here.

$$SARIMA \underbrace{(p,d,q)}_{non-seasonal} \underbrace{(P,D,Q)_m}_{seasonal}$$

Fig. 14. SARIMA Modell

where, m : number of observations each year,
t : auto-regressive element,
d : trend differencing element,
q : trend moving average element,
P,D,Q : same for the seasonal elements

Steps for performing Seasonal ARIMA model :

1) Time Series Visaulization :
It is crucial to visualize the model first and then analyze the time series pattern before building the model.

2) Stationarizing the series :
This is used to check if the series is stationary or not. Dickey-Fuller test is used for testing stationarity.

We have different methods for transforming a non-stationary data to stationary if the p-value in the time series is not significant using the Dickey-Fuller test, just like differencing. Differencing is used for modelling differences of terms than the actual value, denoted by diff() function. For finding the seasonal and ordinal differences(d/D) the functions ndiffs() and nsdiffs() are used.

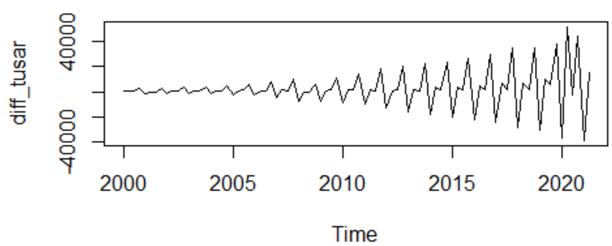

Fig. 15. Stationary Time Series

```
> adf.test(tusa_retail)

        Augmented Dickey-Fuller Test

data:  tusa_retail
Dickey-Fuller = 1.4379, Lag order = 4, p-value = 0.99
alternative hypothesis: stationary

Warning message:
In adf.test(tusa_retail) : p-value greater than printed p-value
> adf.test(diff_tusar)

        Augmented Dickey-Fuller Test

data:  diff_tusar
Dickey-Fuller = -3.0351, Lag order = 4, p-value = 0.1509
alternative hypothesis: stationary

> acf(diff_tusar)
> acf(tusa_retail)
```

Fig. 16. Dickey-Fuller Test

3) Finding Optimal Parameters :
The four parameters p,q,P and Q are gained from the ACF and PACF plots where ACF is the graph for plotting total auto-correlation function and PACF for partial auro-correlation function.

The cut-off can be seen after the first lag in the ACF graph, it can be called as MA(1) process. The dotted blue line gives us an indication of values significantly different from 0.

4) Fitting the ARIMA model :
The ARIMA model is built upon checking various

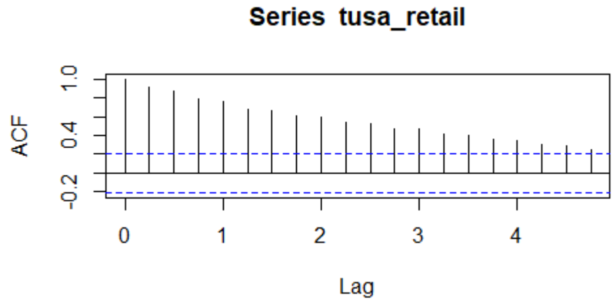

Fig. 17. ACF USA retail

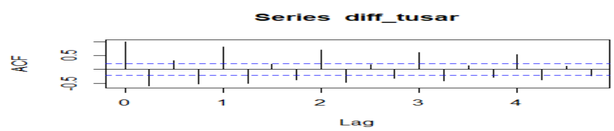

Fig. 18. ACF diff USA retail

combinations of p,d and q. The best model is the one with lowest AICc value. The best fit can be found using the auto.arima() function.

```
> arima_tusar <- auto.arima(diff_tusar)
> plot(arima_tusar)
> arima_tusar
Series: diff_tusar
ARIMA(1,0,0)(1,1,0)[4]

Coefficients:
          ar1     sar1
      -0.3132  -0.6250
s.e.   0.1075   0.1143

sigma^2 estimated as 30766395:  log likelihood=-823.3
AIC=1652.59   AICc=1652.9   BIC=1659.81
```

Fig. 19. auto.arima() function

After doing all the permutations and combinations, the final ARIMA model (2,1,0)(2,1,0)[4] best fits the prediction with AIC value as 1656 and RSME value as 5271.

```
> fit_sarima <- Arima(tusa_retail, order=c(2,1,0), seasonal=list(order=c(2,1,0),period=N
A),
+                     method="ML")
> fit_sarima
Series: tusa_retail
ARIMA(2,1,0)(2,1,0)[4]

Coefficients:
          ar1     ar2     sar1    sar2
      -0.2705  0.1289  -0.6466  0.0117
s.e.   0.1143  0.1156   0.1168  0.1967

sigma^2 estimated as 3.1e+07:  log likelihood=-822.67
AIC=1655.34   AICc=1656.13   BIC=1667.37
```

Fig. 20. auto.arima() function

5) Forecasting and Plotting :

After the final ARIMA is finalised, the values can be predicted and plotted.

```
> summary(fit_sarima)
Series: tusa_retail
ARIMA(2,1,0)(2,1,0)[4]

Coefficients:
          ar1     ar2     sar1    sar2
      -0.2705  0.1289  -0.6466  0.0117
s.e.   0.1143  0.1156   0.1168  0.1967

sigma^2 estimated as 3.1e+07:  log likelihood=-822.67
AIC=1655.34   AICc=1656.13   BIC=1667.37

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 667.0176 5271.874 2378.274 0.2110548 3.716043 0.2395473 -0.01978012
> forecast(fit_sarima, h = 3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       216786.5 209651.2 223921.9 205874.0 227699.1
2021 Q4       255437.7 246605.8 264269.7 241930.4 268945.1
2022 Q1       218137.8 207083.9 229191.6 201232.3 235043.2
```

Fig. 21. Sarima summary

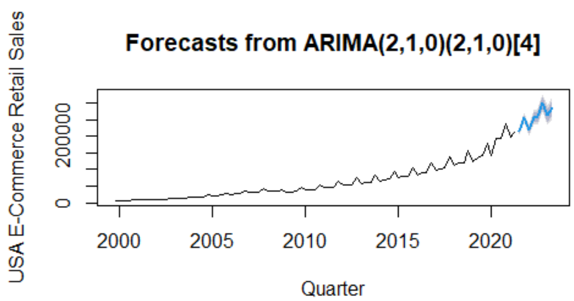

Fig. 22. Sarima Forecast

After comparison of this model with the other models, the SARIMA model has the lowest RSME value as 5271 and hence can be said as the best fit model for predicting the retail sales in the US.

E. Evaluating the Model Fit :

A) Q-Q Plot :
There are few outliers in the below given Q-Q plot with the residuals being normally distributed.

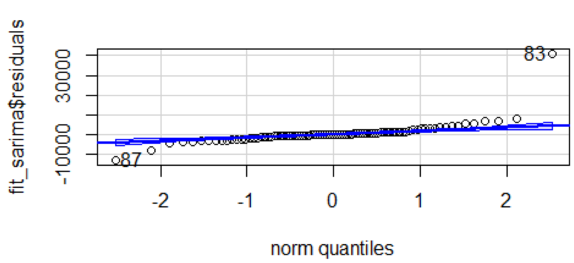

Fig. 23. Q-Q Plot

B) LJung Box Test :
To check whether all the autocorrections are zero is given by the LJung Box test. If the p-value is non-significant, the autocorrelations doesn't differ from 0.

```
> Box.test(fit_sarima$residuals, lag = 1, type = "Ljung")

        Box-Ljung test

data:  fit_sarima$residuals
X-squared = 0.035226, df = 1, p-value = 0.8511
```

Fig. 24. LJung Box Test

C) Checking Residuals :
As per the ACF plot below, it can be seen that the residual auto-correlations do not differ than 0 significantly. Hence, the best fit model. Since, the distribution of residuals is distributed normally and has the constant variance as well, the model best suits for the prediction.

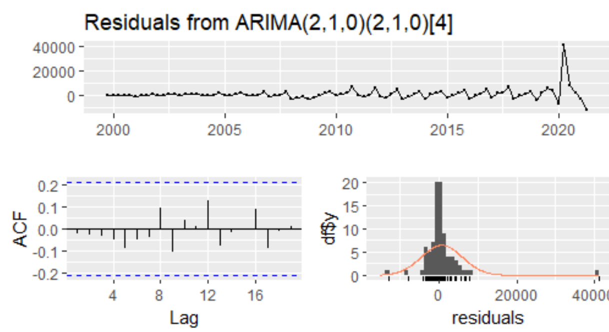

Fig. 25. SARIMA residuals

II. Part B : Logistic Regression

A. Data Description and Data Understanding

Binary Logistic Regression is used for predicting the house categories, Expensive or Budget, of the US region for the chosen period. The prediction is dependent on the various characteristics of the houses such as lotsize, number of bathrooms and bedrooms, waterfront view, fuel, new construction, college education, etc. Whether the house is expensive or is under budget is dependent on various categorical and the numerical variables.

As there are no missing values in the dataset, we are proceeding with the analysis of the of all the necessary factors. The independent variables fuel, waterfront and newConstruction were converted into categorical form from the numerical form. Description of the all the variables that are used in the dataset are given below.

A) Dependent Variable :
1) Name : PriceCat
Type : Dichotomous
Category : Budget(0), Expensive(1)

B) Independent Variable :
1) Name : waterfront
Type : Dichotomous
Category : Yes(1), No(0)

2) Name : newConstruction
Type : Dichotomous
Category : Yes(1), No(0)

3) Name : fuel
Type : Dichotomous
Category : oil(1), electric(2), gas(3)

4) Name : lotsize, age, landValue, livingArea, pctCollege, bedrooms, fireplaces, bathroom, rooms
Type : Continous

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
lotSize	1709	12.20	.00	12.20	.4932	.01615	.66772	.446	7.140	.059	80.729	.118
age	1709	225	0	225	27.76	.699	28.895	834.949	2.522	.059	7.659	.118
landValue	1709	412400	200	412600	34758.35	849.853	35132.968	1234325407	3.092	.059	16.096	.118
livingArea	1709	4612	616	5228	1756.88	14.981	619.308	383542.843	.908	.059	1.298	.118
pctCollege	1709	62	20	82	55.67	.249	10.289	105.868	-1.054	.059	.651	.118
bedrooms	1709	6	1	7	3.15	.020	.813	.661	.353	.059	.468	.118
fireplaces	1709	4	0	4	.60	.013	.556	.309	.398	.059	.743	.118
bathrooms	1709	4.5	.0	4.5	1.905	.0159	.6583	.433	.311	.059	-.438	.118
rooms	1709	10	2	12	7.04	.056	2.316	5.362	.275	.059	-.597	.118
fuel	1709	2	1	3	1.43	.017	.701	.491	1.330	.059	.297	.118
waterfront	1709	1	0	1	.01	.002	.093	.009	10.542	.059	109.265	.118
newConstruction	1709	1	0	1	.05	.005	.211	.045	4.295	.059	16.463	.118
PriceCat	1709	1	1	2	1.45	.012	.498	.248	.182	.059	-1.969	.118
Valid N (listwise)	1709											

Fig. 26. Descriptive Statistics

B. Assumptions

Logistic Regression can be applied to the dichotomous variable for analyzing the set of data. For the application of Logistic Regression, there is a need of meeting all the assumptions that are set for satisfying the criteria of Logistic Regression.

a) Assumption 1 : Dependent variables need to be mutually exclusive
The dependent variable needs to be mutually exclusive, like the observation of the dependent value that have yes as the value, can't have the value of No.

Dependent Variable Encoding

Original Value	Internal Value
Budget	0
Expensive	1

Fig. 27. Dependent Variable Encoding

Classification Table[a,b]

Observed			Predicted: PriceCat: Budget	Predicted: PriceCat: Expensive	Percentage Correct
Step 0	PriceCat	Budget	932	0	100.0
		Expensive	777	0	.0
	Overall Percentage				54.5

a. Constant is included in the model.
b. The cut value is .500

Fig. 28. Classification Table

From the above table it can seen that the Budget is categorized as 0 and expensive is classified as 1. There are 932 values in the Budget price category valued as 0 and 777 observations in the expensive category valued as 1. Both are the mutually exclusive events. Thus, meeting the criteria for Logistic Regression.

c) Assumption 2 : Sample size
The size of the sample needs to be large enough for logistic regression to work. A small sample size with many independent variables gives us the inappropriate results. This model has total of 1709 observations, good sample size for applying logistic regression to the model.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1709	100.0
	Missing Cases	0	.0
	Total	1709	100.0
Unselected Cases		0	.0
Total		1709	100.0

a. If weight is in effect, see classification table for the total number of cases.

Fig. 29. Case processing summary

c) Assumption 3 : Absence of Collinearity
When two or more independent variables are related to each other, multicollinearity occurs. This gives a problem in determining the logistics regression model because it becomes difficult in determining which independent variable contributes to the variance of the dependent variable.

There are two methods for determining the multicollinearity :

1) If the coefficient value in the correlation matrix are not in between -0.7 and 0.7, them the predictors are considered as multicollinear.
2) VIF value to be less than 10.

Since, the values in the table above doesn't lie between -0,7 and 0.7, the variables aren't multicollinear.

Correlation Matrix

		Constant	fuel	waterfront	newConstruction	lotSize	age	landValue	livingArea	pctCollege	bedrooms	fireplaces	bathrooms	rooms
Step 1	Constant	1.000	-.426	-.112	-.016	-.030	-.071	-.027	-.095	-.652	-.209	.131	-.286	-.015
	fuel	-.426	1.000	-.091	.064	-.297	-.186	.172	-.044	.135	.083	.080	.072	.006
	waterfront	-.112	-.091	1.000	-.015	.045	-.041	-.026	.012	.081	.084	-.036	.021	.000
	newConstruction	-.016	.064	-.015	1.000	-.005	.128	-.027	-.115	.074	.015	.093	-.051	-.005
	lotSize	-.030	-.297	.045	-.005	1.000	.066	-.003	-.035	.052	-.070	-.054	.071	-.010
	age	-.071	-.186	-.041	.128	.066	1.000	-.254	-.066	.014	-.154	.079	.313	-.016
	landValue	-.027	.172	-.026	-.027	-.003	-.254	1.000	.121	-.353	-.060	.055	.091	-.026
	livingArea	-.095	-.044	.012	-.115	-.035	-.066	.121	1.000	-.074	-.301	-.165	-.310	-.374
	pctCollege	-.652	.135	.081	.074	.052	.014	-.353	-.074	1.000	-.001	-.169	-.021	.001
	bedrooms	-.209	.083	.084	.015	-.070	-.154	-.060	-.301	-.001	1.000	-.004	-.059	-.363
	fireplaces	.131	.080	-.036	.093	-.054	.079	.055	-.165	-.169	-.004	1.000	-.132	.008
	bathrooms	-.286	.072	.021	-.051	.071	.313	.091	-.310	-.021	-.059	-.132	1.000	-.008
	rooms	-.015	.006	.000	-.005	-.010	-.016	-.026	-.374	.001	-.363	.008	-.008	1.000

Fig. 30. Correlation Matrix

d) Assumption 4 : Independence Of errors

The assumption says that the residual values must stand independent, no error terms should be connected to each other. Durbin-Watson statistics is used for verification of the assumption.

The auto-correlation amongst the residual values is verified by the Durbin-Watson statistics, it ranges from 1 to 3. A good statistics is when the value lies close to 2. In the above table, we get the value of 1.702, which has less auto-correlation amongst the residuals.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.265[a]	.070	.069	.481	1.702

a. Predictors: (Constant), newConstruction, waterfront, fuel

b. Dependent Variable: PriceCat

Fig. 31. Durbin-Watson Statistics

e) Assumption 5 : No significant Outliers

For any model, highly influential points also known as the high leverage points, should not be present because it can bias the predicted performance. This can cause discrepancies in the predicted model and the overall predictions. This can be calculated using the Cooke's distance in SPSS. The Cooke's distance should be less than 1. The point where the Cooke's distance exceeds 1 is considered as the outlier.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.87	3.06	1.45	.330	1709
Std. Predicted Value	-1.781	4.875	.000	1.000	1709
Standard Error of Predicted Value	.015	.166	.030	.013	1709
Adjusted Predicted Value	.86	3.23	1.45	.332	1709
Residual	-1.065	1.101	.000	.373	1709
Std. Residual	-2.846	2.943	.000	.996	1709
Stud. Residual	-3.054	2.966	.000	1.001	1709
Deleted Residual	-1.227	1.118	.000	.376	1709
Stud. Deleted Residual	-3.062	2.973	.000	1.002	1709
Mahal. Distance	1.742	333.389	11.993	17.318	1709
Cook's Distance	.000	.109	.001	.004	1709
Centered Leverage Value	.001	.195	.007	.010	1709

a. Dependent Variable: PriceCat

Fig. 32. Residual Statistics

The minimum and maximum Cooke's distance is 0.000 and 0.109 respectively. Hence, the dataset doesn't consist of highly influential points.

C. Understanding and Building a Model

While building a Logistic Regression model, two different outputs are obtained.

Block 0 :

The null model known as the Block 0, doesn't consist any independent variable. Here, house category is predicted without any independent variable. Later, the accuracy is built by adding the independent variable to the model.

Block 1 :

Using the 0.05 significance level, a global hypothesis test is conducted for checking for the regression coefficients other than 0. The null hypothesis has coefficients of independent variables as 0. The Chi-square value is taken into consideration for step, block and model.

a) Model I :

For building the first model, priceCategory which has the two components, Expensive and Budget, is considered as the independent variable with all the rest 13 variables as the dependent ones. The p-value for each of the dependent variable is given in the table below.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lotSize	.542	.146	13.840	1	.000	1.719
	age	-.006	.003	3.168	1	.075	.994
	landValue	.000	.000	115.218	1	.000	1.000
	livingArea	.002	.000	86.115	1	.000	1.002
	pctCollege	-.015	.008	3.599	1	.058	.985
	bedrooms	-.109	.129	.709	1	.400	.897
	fireplaces	.047	.149	.098	1	.754	1.048
	bathrooms	1.002	.161	38.904	1	.000	2.724
	rooms	.016	.048	.104	1	.747	1.016
	fuel	-.124	.118	1.108	1	.293	.883
	waterfront	3.426	.956	12.839	1	.000	30.742
	newConstruction	-.245	.443	.306	1	.580	.783
	Constant	-6.526	.598	119.134	1	.000	.001

a. Variable(s) entered on step 1: lotSize, age, landValue, livingArea, pctCollege, bedrooms, fireplaces, bathrooms, rooms, fuel, waterfront, newConstruction.

Fig. 33. Variables in the equation

When the Omnibus test is run for the model, the Chi-square value for the step, block and model comes out to be 1071.357, which is a bit high value with the p-value as 0.001. Hosmer and Lemeshow Test gov es the Chi-square value as 25.9 with 0.001 as the p-value for the model. It has the overall accuracy of 83.4%

The classification table below gives us the specificity as 88.2% and sensitivity as 77.6% for the model for the cut value 0.05. It has the overall accuracy of 83.4%. But since there are non-significant parameters considered in the model, we reject this for better accuracy.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1071.357	12	.000
	Block	1071.357	12	.000
	Model	1071.357	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1283.743[a]	.466	.623

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	25.964	8	.001

Fig. 34. Model Summary

Classification Table[a]

Observed			Predicted PriceCat Budget	Predicted PriceCat Expensive	Percentage Correct
Step 1	PriceCat	Budget	822	110	88.2
		Expensive	174	603	77.6
	Overall Percentage				83.4

a. The cut value is .500

Fig. 35. Classification Table

b) Model II :
In this model, the significant non-significant parameters are dropped out; bedrooms, fireplaces, rooms and newConstruction, to yield better accuracy. The independent variables used in these models are lotsize, landvalue, livingarea, bathroom, waterfront, age, pctcollege and fuel.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lotSize	.536	.145	13.722	1	.000	1.709
	landValue	.000	.000	115.240	1	.000	1.000
	livingArea	.002	.000	135.429	1	.000	1.002
	bathrooms	.998	.158	39.634	1	.000	2.712
	waterfront	3.500	.950	13.581	1	.000	33.105
	age	-.006	.003	3.702	1	.054	.994
	pctCollege	-.015	.008	3.356	1	.067	.985
	fuel	-.115	.116	.980	1	.322	.891
	Constant	-6.669	.575	134.309	1	.000	.001

a. Variable(s) entered on step 1: lotSize, landValue, livingArea, bathrooms, waterfront, age, pctCollege, fuel.

Fig. 36. Variables used in the model

The cut value with 0.4 has specificity as 82.7% and sensitivity as 82.5% with the overall accuracy as 82.6%.
The cut value with 0.5 has specificity as 87.8% and sensitivity as 77.6% with the overall accuracy as 83.1%.
The cut value with 0.6 has specificity as 90.2% and sensitivity as 72.1% with the overall accuracy as 82.0%.

Classification Table[a]

Observed			Predicted PriceCat Budget	Predicted PriceCat Expensive	Percentage Correct
Step 1	PriceCat	Budget	771	161	82.7
		Expensive	136	641	82.5
	Overall Percentage				82.6

a. The cut value is .400

Fig. 37. Cut Value as 0.4

Classification Table[a]

Observed			Predicted PriceCat Budget	Predicted PriceCat Expensive	Percentage Correct
Step 1	PriceCat	Budget	818	114	87.8
		Expensive	174	603	77.6
	Overall Percentage				83.1

a. The cut value is .500

Fig. 38. Cut Value as 0.5

Classification Table[a]

Observed			Predicted PriceCat Budget	Predicted PriceCat Expensive	Percentage Correct
Step 1	PriceCat	Budget	841	91	90.2
		Expensive	217	560	72.1
	Overall Percentage				82.0

a. The cut value is .600

Fig. 39. Cut Value as 0.6

From the accuracy above for three different cut values, the cut value with 0.5 stands the most optimum one. Since, the model still has three non-significant parameters; age, pctCollege and fuel, the model is rejected for better accuracy.

c) Final Model :
For building the final model, all the parameters with p-value as 0.000 are considered with some applied transformations on the independent variables. The parameters taken into consideration are lotsize, bathrooms, waterfront, log transformation of the livingArea and squareroot of price.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lotSize	.519	.140	13.730	1	.000	1.681
	bathrooms	1.127	.146	59.867	1	.000	3.087
	waterfront	3.764	.972	15.002	1	.000	43.099
	log_livingarea	8.597	.773	123.813	1	.000	5415.649
	sqrt_price	.014	.001	134.113	1	.000	1.014
	Constant	-32.634	2.378	188.291	1	.000	.000

a. Variable(s) entered on step 1: lotSize, bathrooms, waterfront, log_livingarea, sqrt_price.

Fig. 40. Variables in the Final Model

When the Omnibus test is run for the model, the Chi-square value for the step, block and model comes out to be 1042.98

with the p-value as 0.000. Hosmer and Lemeshow Test gov es the Chi-square value as 14.25 with 0.075 as the p-value for the model.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1042.988	5	.000
	Block	1042.988	5	.000
	Model	1042.988	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1312.112[a]	.457	.611

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	14.251	8	.075

Fig. 41. Model Summary

Classification Table[a]

Observed			Predicted: PriceCat Budget	Predicted: PriceCat Expensive	Percentage Correct
Step 1	PriceCat	Budget	814	118	87.3
		Expensive	178	599	77.1
	Overall Percentage				82.7

a. The cut value is .500

Fig. 42. Classification Table

The classification table below gives us the specificity as 87.3% and sensitivity as 77.1% for the model for the cut value 0.05. The accuracy of the final model comes out to be 82.7%, which is the best fit for the model when specificity, sensisitivity, accuracy and p-value are taken into consideration.

III. Conclusion

Time series concludes after the comparison of Seasonal ARIMA model with the other models, the SARIMA model has the lowest RSME value as 5271 and hence can be said as the best fit model for predicting the retail sales in the US. In case of the Logistic Regression, the accuracy of the final model comes out to be 82.7%, which is the best fit for the model when specificity, sensisitivity, accuracy and p-value are taken into consideration.

References

[1] G. Nunnari and V. Nunnari, ”Forecasting Monthly Sales Retail Time Series: A Case Study,” 2017 IEEE 19th Conference on Business Informatics (CBI), 2017, pp. 1-6, doi: 10.1109/CBI.2017.57.

[2] P. Durganjali and M. V. Pujitha, ”House Resale Price Prediction Using Classification Algorithms,” 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.