Time Series, Binary Logistic Regression, and Principal Component Analysis

Nivedita Vishwanath Hiremath
MSc in Data Analytics
x21108471@student.ncirl.ie

***Abstract*— In this paper, an attempt is made to use SPSS and R to perform Time Series analysis, Binary Logistic Regression, and Principal Component Analysis on the relevant datasets.**

I. Objective Of Time Series Analysis

The goal of Time series analysis is to identify the kind of time series in a dataset and then apply the relevant time series models to the data. It's also required to consider certain comparisons. The values of the Akaike Information Criterion(AIC) and the Root Mean Square Error(RMSE) are considered. Forecasting will be done using the best model available.

A. About dataset and description

The quarterly dataset is of United States e-commerce retail sales is in CSV format from fourth-quarter 1999 to second quarter 2021.

B. Time Series Analysis and model building steps

Identifying Pattern -The first step in time series analysis is to identify the pattern in the data. By looking at Fig .1. The pattern seems to be the trend with the linear increase over time. But to confirm this hold seasonality pattern and to analyze any changes in seasonality over time, season plot Fig .2. and season subseries Fig .3. the graph was plotted.

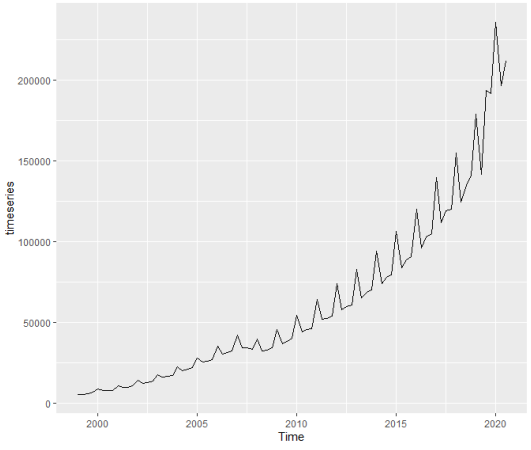

Fig .1. Time series pattern

Fig .2. is the seasonal plot that is plotted against data and individual quarters by observing this plot we can say that in Quarter 1 sales in high and it dropped in Quarter 2 and gradually increased in Quarter 3, again dropping in Quarter 4. Fig .3. is an alternative seasonal plot that shows changes in seasonality over time and it also indicates the means of each quarter, the mean of Quarter 1 is high. Thus, we can say that the data is having both trend and seasonality.

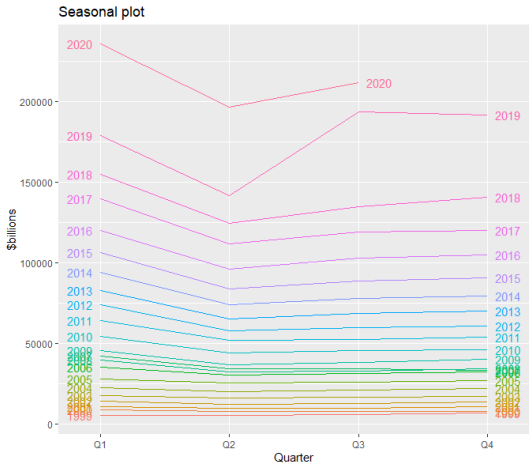

Fig . 2. Quarter-wise sales

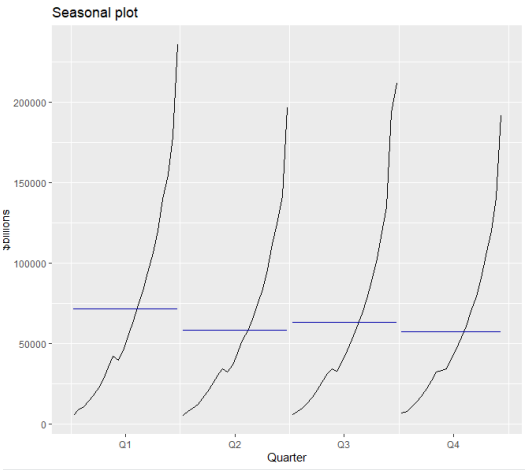

Fig .3. Seasonal plot Quarter

Seasonal Decomposition - The Time-series quarterly data has a seasonal aspect to it, and it can be decomposed into a trend component(which captures in level over time), a seasonal component(which captures effects due to the time of year), and an irregular component(which captures those influences not addressed by trend or

seasonal)[1].By looking at above Fig .1. and Fig .2. Additive decomposition is suitable because the magnitude of seasonal fluctuations in earlier years is similar to the size of seasonal fluctuations in later periods. It is also used to estimate trend, seasonality, and irregularity effect on time t.

The additive model is calculated by $Y_t = Trend_t + Seasonal_t + Irregular_t$(where t is a time of observation).Fig .4. shows additive decomposition plot.

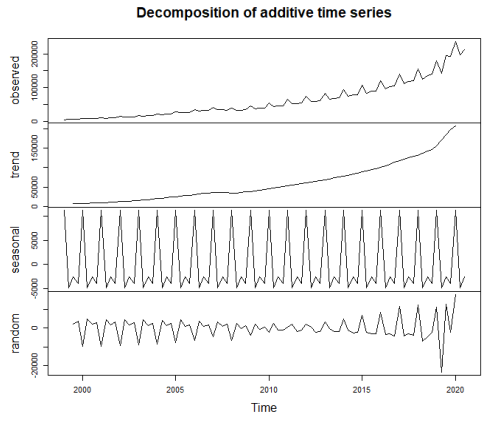

Fig .4. Additive decomposed time series

C. Model Building and Forecast Accuracy

Model 1: There are several simple forecasting methods, but we chose the seasonal naïve method because our time series data has seasonal and upward trend data[2] and by observing Fig .2. for the Quarterly data available, the forecast for the Quarter1 of a year is the same as the observed number for the same quarter the previous year, it follows a similar pattern for other Quarters. The graph Fig .5. was plotted for the forecast of period 3 and the RMSE value obtained 15144 from Fig .6. was noted.

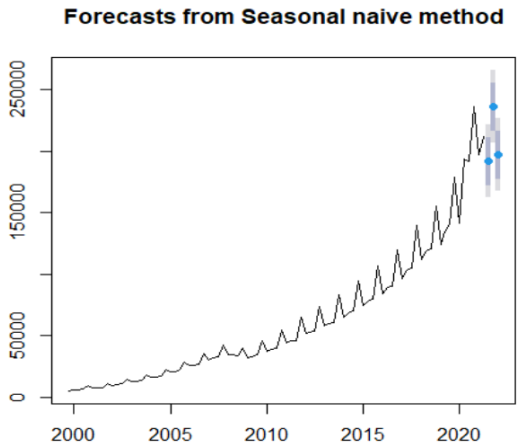

Fig .5. Forecast using Seasonal naïve method

```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = timeseries, 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 .6. Seasonal naïve output properties

Model 2: Exponential smoothening model. A triple exponential model known as Holt-Winters exponential smoothing is an extension of Holt exponential smoothening chosen because it fits the time series which has level, trend, and seasonal components in it[3]. Here more weight is given to recent observations, and weight decreases exponentially as time passes. There are two variations to this method with a respective seasonal component named 'Additive' and 'Multiplicative' methods. It is a bit difficult to say whether the seasonality is Additive or Multiplicative. We have applied both the methods to our time series data and AIC, RMSE values were noted. From Fig .7. the AIC value for Holt-Winters 'Additive' method obtained was 1916.303 and from Fig .8. the AIC value for Holt-Winters 'Multiplicative' method obtained was 1768.792. The respective RMSE values were 5865.446 and 4969.59.

```
Holt-Winters' additive method

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

  Smoothing parameters:
    alpha = 0.5904
    beta  = 0.0769
    gamma = 0.4096

  Initial states:
    l = 2465.0902
    b = 831.4089
    s = -3867.664 -2172.807 -4706.572 10747.04

  sigma:  6155.27

     AIC     AICc      BIC
1916.303 1918.641 1938.496
```

Fig .7. Holt-Winters Additive method

```
Holt-Winters' multiplicative method

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

  Smoothing parameters:
    alpha = 0.4666
    beta  = 0.1178
    gamma = 1e-04

  Initial states:
    l = 4702.8499
    b = 530.3473
    s = 0.949 0.9662 0.9074 1.1775

  sigma:  0.0641

     AIC     AICc      BIC
1768.792 1771.130 1790.985
```

Fig .8. Holt-Winters Multiplicative method

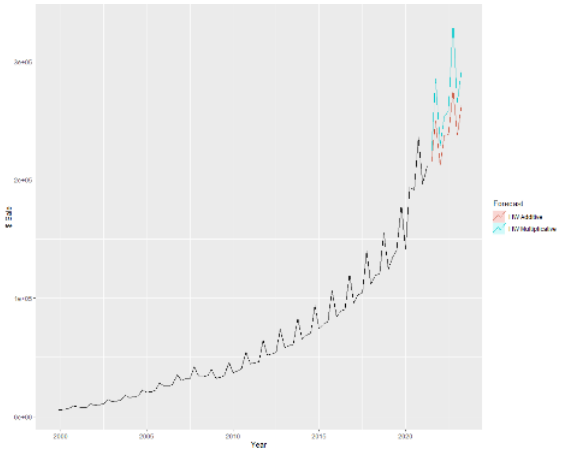

Fig .9. Holt-Winters model plot for both methods

When compared RMSE values from both the models. The method with 'Multiplicative seasonality' gave the least RMSE and AIC value, this was expected because from Fig .9. it depicts seasonal variation in data increased as the level of the series increased. Hence Holt-Winters multiplicative seasonality fits the data best.

```
> forecast(hwFit2,h=3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       220764.5 202621.2 238907.8 193016.7 248512.3
2021 Q4       285631.2 258708.3 312554.1 244456.1 326806.2
2022 Q1       229140.2 204320.5 253959.8 191181.8 267098.6
```

Fig .10. Forecast for Holt-Winters Multiplicative method

Model 3: There is one more method called ETS exponential model selection. ETS stands for error type, trend type, seasonal type. We have used the ZZZ parameter in a function, which lets the function decide the optimum result with the best parameters.

```
ETS(M,A,M)

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

  Smoothing parameters:
    alpha = 0.6684
    beta  = 0.0577
    gamma = 0.2913

  Initial states:
    l = 4209.3873
    b = 610.6294
    s = 0.9482 0.9422 0.9674 1.1422

  sigma:  0.0505

     AIC     AICc      BIC
1726.852 1729.189 1749.045
```

Fig .11. ETS function.

The ETS model has taken M, A, M(multiplicative error type, additive trend, multiplicative seasonality) parameters as the best-fitting model for the data. Earlier also we saw that our time series data has an Additive trend and Multiplicative seasonality which is fair enough. The RMSE value obtained was 5397.091.

```
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       205012.7 191750.0 218275.4 184729.2 225296.2
2021 Q4       257797.1 237362.6 278231.5 226545.2 289048.9
2022 Q1       209989.7 190504.3 229475.1 180189.3 239790.1
```

Fig .12. Forecast for ETS

Model 4: Next we took a popular time series forecast model known as ARIMA. We have considered SARIMA as in our time series data we have a seasonality component.

SARIMA is denoted as –

SARIMA (p, d, q) $(P, D, Q)_m$

Where m is number of observation per year and p ,d ,q are the non seasonal part of the model. P,D,Q are the seasonal part of the model[4]

As our data has trend in it we have to render it as stationary, we can do it by differencing. There are two tests to check stationarity those are Augmented Dickey- Fuller test and KPSS. In R adf.test() package is used to check stationarity for the Augmented Dickey-Fuller test and ndiffs() is used for KPSS to check stationarity. The adf.test() Fig .13. gave p-value >0.05 which says non-stationary and ndiffs() gave value as 1 which means 1 number of times time series must be differentiated to make it stationary.

```
> adf.test(timeseries)

        Augmented Dickey-Fuller Test

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

Fig .13. Augmented Dickey-Fuller Test

```
> ndiffs(timeseries)
[1] 1
```

Fig .14. ndiffs function

We have used diff() function in R to make it stationary. The next step is to determine p, d, q, and P, D, Q values which can be done by analyzing ACF and PACF graphs. But, we have used auto.arima() function in R to get p, d, q and P,D,Q values.

```
Series: d
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 .15. Auto Arima function properties

The optimal suggestion given by function was ARIMA(1,0,0)(1,1,0) spike at lag 4 here d=0

because we have already differentiated it to make non-seasonal part to stationary. The same values are given to the model and obtained RMSE 5349.757 value was noted.

```
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

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE         ACF1
Training set 793.2327 5349.757 2418.634 269.9636 300.9593 0.9515099 0.006252549
```

Fig .16. Sarima Model values

To check whether our model is proper or not the residuals and ACF graph was plotted Fig .17. and "Ljung-Box" test was conducted Fig .18.

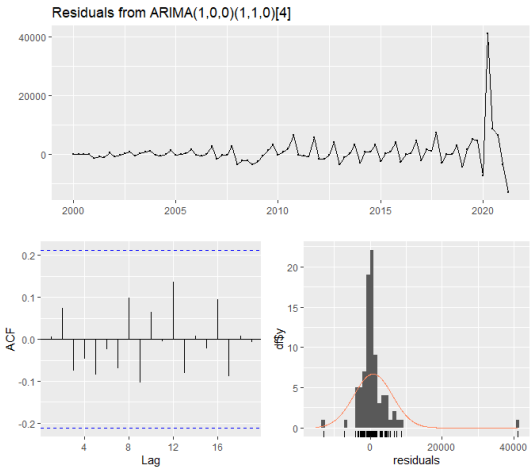

Fig .17. Checking residuals plot

```
> Box.test(sarima$residuals,type="Ljung-Box")

        Box-Ljung test

data:  sarima$residuals
X-squared = 0.0034808, df = 1, p-value = 0.953
```

Fig .18. Ljung-Box test

From Fig .17. The residuals are centered around zero and from the ACF plot for all lags, the autocorrelations are zero, i.e. the residuals are normally and independently distributed (no association between them).

From Fig .18. the p-value is 0.953 implying that the autocorrelations aren't significantly different from zero.

```
> forecast(sarima,h=3)
        Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
2021 Q3       6346.746  -761.6955  13455.19  -4524.677  17218.17
2021 Q4      39510.064 32061.2308  46958.90  28118.057  50902.07
2022 Q1     -37680.064 -45161.4434 -30198.68 -49121.846 -26238.28
```

Fig .19. Forecast of Sarima model

D. Results and Interpretation

As per our analysis, our data holds trend, seasonality, and irregularity, few models were chosen for accuracy prediction from different categories. So, amongst different simple models, Seasonal naïve was chosen for accounting seasonality. From Exponential smoothening Holt-Winters, seasonality model was chosen for accounting level, trend, and a seasonal component and from the same exponential smoothening category ETS was chosen for accounting error, trend, and seasonality. Finally, the Seasonal ARIMA model known as SARIMA was chosen.

All four models are evaluated with the common factor RMSE. Results are in the table below-

TABLE I Results with RMSE values

Model	***RMSE***
Seasonal Naive	15144
Holt-Winters('multiplicative')	4969.59
ETS	5397.09
SARIMA	5349.757

By looking at TABLE I we can conclude that Holt-Winter('multiplicative') has the least RMSE and is more accurate than other models since it gives more weightage to recent observations, hence it is a good model for forecasting.

From Fig .8. the smoothening parameters and its values of coefficients are alpha (level smoothening) = 0.4666, beta (trend smoothening)=0.1178, and gamma(seasonal smoothening)=1e-04. The gamma parameter is least this depicts seasonal component does not alter much over time. From Fig .10. for Holts-Winter the point forecast shows the time series for the next three quarters value and Lo 80 and Hi 80 lower and upper bound of 80% confidence interval. Similarly, Lo 95 and Hi 95 is a lower and upper bound of 95% confidence interval. This confidence interval states that the predicted value will fall within the stated range.

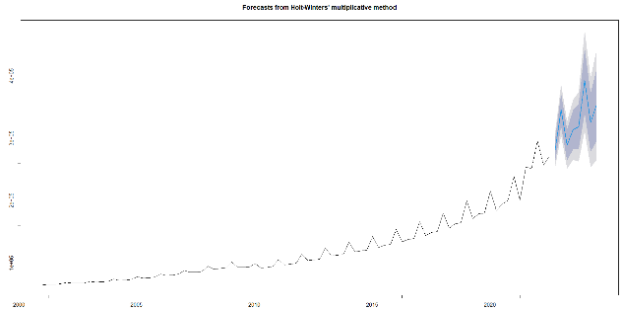

Fig .20. Holts-Winter Forecast

From above Fig .20. Forecast values for three periods ahead can be seen visually. The black-colored line shows original time series data and the blue-colored line is the forecasted value appended

to original data. There can be seen two-colored area shaded grey and blue. These are the prediction interval region that we discussed above. The blue line is with 80% confidence interval and the grey shaded area is with 95% confidence interval.

II. Objective Of Binary Logistic Regression

The goal of the analysis is to perform Binary logistic regression to determine binary outcomes with different independent variables. To compute the probability of a response variable, logistic regression employs a separate function known as the 'Logit Function' (Inverse of Sigmoid function).

$$\mathbf{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

p = probability

$\frac{p}{1-p}$ = corresponding odds

All the important measures and assumptions are taken to build a model and verify the goodness of fit.

A. Dataset Description, Assumptions, and Analysis

The dataset consists of characteristics of houses, sold in the US region. It consists of 1709 rows. The dependent variable is 'PriceCat' Fig .21. where Budget is encoded as 0 and Expensive is encoded as 1.

Dependent Variable Encoding

Original Value	Internal Value
Budget	0
Expensive	1

Fig .21. Dependent variable encoding

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			PriceCat		Percentage Correct
	Observed		Budget	Expensive	
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 .22. Block 0 -Null Model

Block 0- Null Model Fig .22. is the analysis without any of the independent variables, this is a baseline comparing the model. So, by adding independent variables in our model the overall percentage should be greater than 54.5%.

The Goodness-of-fit statistics to describe Model

Omnibus Test –If this test result shows significant value, then we can say the model is fit and there is an improvement in fit compared to the null model.

Hosmer and Lemeshow Test – The model is a good fit if the significance value is greater than 0.05. If it is less than 0.05 then it describes as no difference between the observed and predicted model.

Nagelkerke R Square – This covers the range from 0 to 1 and is used as an approximate variation in the criterion variable.

Deviance(-2 Log likelihood) – Lower the deviance value the better the model fit, it is the sum of squares of residuals.

B. Model building steps

From Fig .23. the correlation matrix we can see all variables are significant and good to use in model prediction. We also have categorical variables in our dataset namely fuel, waterfront, new construction, rooms, bathrooms, bedrooms, fireplaces. Since the equation only handles binary and continuous types, these variables need to convert dummy variables it is very easy to do in SPSS Fig .24. By providing columns to Categorical Covariates corresponding dummy values will be generated.

Correlations

		lotSize	age	landValue	livingArea	pctCollege	bedrooms	fireplaces	bathrooms	rooms	PriceCat
lotSize	Pearson Correlation	1	-.013	.069**	.179**	-.025	.121**	.099**	.099**	.148**	.168**
	Sig. (2-tailed)		.603	.005	.000	.309	.000	.000	.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
age	Pearson Correlation	-.013	1	-.017	-.179**	-.037	.018	-.173**	-.362**	-.086**	-.154**
	Sig. (2-tailed)	.603		.483	.000	.130	.457	.000	.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
landValue	Pearson Correlation	.069**	-.017	1	.424**	.225**	.206**	.209**	.295**	.300**	.414**
	Sig. (2-tailed)	.005	.483		.000	.000	.000	.000	.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
livingArea	Pearson Correlation	.179**	-.179**	.424**	1	.205**	.656**	.474**	.721**	.733**	.667**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
pctCollege	Pearson Correlation	-.025	-.037	.225**	.205**	1	.165**	.247**	.171**	.157**	.173**
	Sig. (2-tailed)	.309	.130	.000	.000		.000	.000	.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
bedrooms	Pearson Correlation	.121**	.018	.206**	.656**	.165**	1	.287**	.466**	.670**	.409**
	Sig. (2-tailed)	.000	.457	.000	.000	.000		.000	.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
fireplaces	Pearson Correlation	.099**	-.173**	.209**	.474**	.247**	.287**	1	.438**	.320**	.322**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
bathrooms	Pearson Correlation	.099**	-.362**	.295**	.721**	.171**	.466**	.438**	1	.521**	.551**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000	.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
rooms	Pearson Correlation	.148**	-.086**	.300**	.733**	.157**	.670**	.320**	.521**	1	.466**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000		.000
	N	1709	1709	1709	1709	1709	1709	1709	1709	1709	1709
PriceCat	Pearson Correlation	.168**	-.154**	.414**	.667**	.173**	.409**	.322**	.551**	.466**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	

Fig .23. Correlation matrix

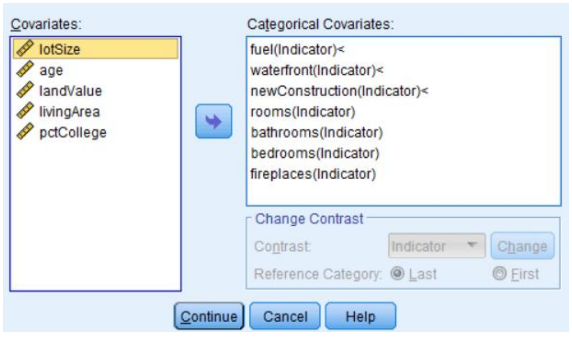

Fig .24. Creating Dummy Columns in SPSS

Model 1: Initially we took all predictor variables for checking the probability of the dependent variable with a classification threshold of 0.5. Although Omnibus test of model coefficients showed significant value but failed in Hosmer and Lemeshow Test by giving p-value <0.05 Fig .25.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	21.859	8	.005

Fig .25. Model 1 Hosmer and Lemeshow Test

Model 2: Removing few of the variables which showed insignificance value in Wald statistic. The variables lotsize, age, pctcollege, fireplaces, bathrooms are removed. The classification threshold considered here was 0.5. The Omnibus test showed significant value and Hosmer and Lemeshow Test gave non-significant value of 0.198. Since both the test have passed, we have calculated the value of -2 log likelihood and Nagaelkerke R square. 1333.975 and 0.601 are the respective values Fig .26. Overall accuracy classification percentage is 82.1%.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1021.125	22	.000
	Block	1021.125	22	.000
	Model	1021.125	22	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1333.975[a]	.450	.601

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.071	8	.198

Fig .26. Model 2 Summary results

Model 3: Since model 2 passed with Omnibus test and Hosmer and Lemeshow Test. We have taken the same variables as predictors with a different threshold value of 0.4 just to check whether it increases the overall accuracy classification percentage. In this model, the Omnibus test showed significant value. Hosmer and Lemeshow test gave non-significant value of 0.198. There was no improvement in -2 log likelihood value which was obtained 1333.975 and slightly there was an increase in overall accuracy classification percentage value of 0.5% the obtained value was 82.5% Fig .27.

Classification Table[a]

			Predicted		
			PriceCat		Percentage Correct
	Observed		Budget	Expensive	
Step 1	PriceCat	Budget	752	180	80.7
		Expensive	119	658	84.7
	Overall Percentage				82.5

a. The cut value is .400

Fig .27. Model 3 Classification Table

Model 4: In this model, we have tried to improve, deviance(-2 log likelihood) to get the least value. The independent variables considered here are variables lotsize, landvalue, livingarea, bedrooms, bathrooms, rooms, waterfront and interaction between landvalue*livingArea*lotsize variables. The Omnibus showed the significant result. Hosmer and Lemeshow Test showed non-significant value of 0.102. Since we are using the interaction between variables it is an important factor to consider significance value, as we got 0.000 as significance, it depicts non-linearity between independent variable and logit, it is good to consider. There was a significant drop of deviance(-2 log likelihood) compared to previous models, the value obtained was 1246.424 and Nagelkerke R Square value was 0.638 Fig .28. The Overall Classification percentage accuracy obtained was 83%.

So far this is good model with least deviance -2 log likelihood value and with so far percentage accuracy classification(PAC) value.

Below Fig .28. shows Model 4 summary.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1108.675	29	.000
	Block	1108.675	29	.000
	Model	1108.675	29	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1246.424[a]	.477	.638

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.282	8	.102

Fig .28. Model 4 Summary

Model 5: In this model we tried to implement one of the dimension reduction techniques know as PCA(Principal Component Analysis) this helps in transforming a large number of correlated variables to smaller uncorrelated variables. This technique comes up with super variables combining different independent variables they named as PC1, PC2, PC3, and so on. These components are the linear combinations of original variables.

To apply this test sample size should be 10-20 observations for each variable. Our dataset satisfies this assumption. The second step is to check correlation between the variables and there has to be a present certain correlation between variables i.e > 0.3. Fig .23. correlation plot shows we have a certain correlation between the variables in our dataset. The third step is to check KMO and Bartlett's Test, to meet this assumption Bartlett's Test of Sphericity should be significant i.e < 0.05, and Kaiser-Meyer-Olkin (KMO) this value ranges from 0 to 1 and it measures the sampling adequacy. By conducting these tests Fig .29. the assumption is satisfied.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.788
Bartlett's Test of Sphericity	Approx. Chi-Square	5968.746
	df	66
	Sig.	.000

Fig .29. KMO and Bartlett's Test

To apply PCA in SPSS we have applied a transformation to variables fuel, waterfront, and newconstruction to numerical as they were string in the dataset, for fuel the value 'oil':1, 'electric':2, 'gas':3, for waterfront the value 'No':0, 'Yes':1 and for newconstruction 'No':0, 'Yes':1. To select the number of components the eigenvalue was chosen was greater than 1, it tells how much variance is accounted by that particular component and scree plot was also plotted to choose components.

From Fig .30. explains that there are total of 12 components generated out of that component 1 has the highest percentage variation explanation of 30.843% and the eigenvalue obtained was 3.701. Out of 12 components, 5 components are showing eigenvalue > 1. We considered only those 5 components in our model building because they simplified and cumulatively gave 70% of the variation of information.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.701	30.843	30.843	3.701	30.843	30.843	3.268	27.236	27.236
2	1.370	11.418	42.262	1.370	11.418	42.262	1.468	12.234	39.470
3	1.217	10.142	52.404	1.217	10.142	52.404	1.312	10.935	50.404
4	1.093	9.109	61.513	1.093	9.109	61.513	1.213	10.112	60.516
5	1.026	8.552	70.065	1.026	8.552	70.065	1.146	9.549	70.065
6	.875	7.289	77.354						
7	.661	5.505	82.859						
8	.631	5.256	88.115						
9	.540	4.497	92.612						
10	.395	3.289	95.901						
11	.313	2.611	98.512						
12	.179	1.488	100.000						

Extraction Method: Principal Component Analysis.

Fig .30. Total variation of components

Scree plot is the plot of eigenvalues with the components in the x-axis. By looking at Fig .31. we differentiated which are important components and which are not so important. To cut off, we looked at the elbow line in the plot. The elbow point we found at 6 and we considered all the components above that point.

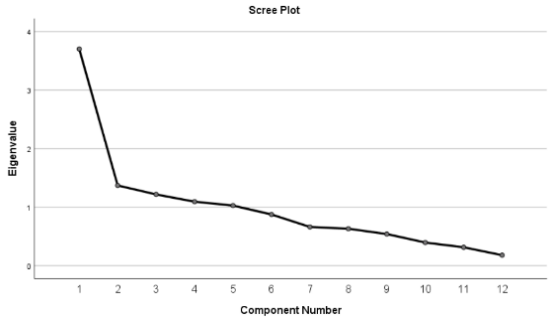

Fig .31.Scree plot

Rotated component Fig .32. helps in analyzing the correlation between each component and all the

underlying variables. We can even label the components as well.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
newConstruction	.326	.270	.261	-.624	
fuel		.250	.713		
waterfront					.853
rooms	.865				
bathrooms	.658	.510			
fireplaces	.408	.415		.448	
bedrooms	.849				
pctCollege			.255	.754	
livingArea	.878				
landValue	.426				.588
age		-.902			
lotSize			-.762		

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 6 iterations.

Fig .32. Rotated Component.

The selected 5 components were applied to Binary logistic regression with a classification cutoff threshold of 0.5. The results obtained were Fig .29.the omnibus of a model coefficient is satisfied and Hosmer and Lemeshow test also passed with a value of 0.127 non-significant.But compared to all the above models the deviance(-2 Log likelihood) is high. The value obtained was 1407.708 Fig .33

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	947.392	5	.000
	Block	947.392	5	.000
	Model	947.392	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1407.708[a]	.426	.569

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.571	8	.127

Fig .33. Model 5 Summary

C. Results and Interpretation

We have applied several models to classify the house price category, by considering the fact of selecting significant variables, changing cutoff threshold, getting least deviance(-2 Log likelihood),applying dimension reduction.

Model 1 was rejected because it showed a significant p-value for Hosmer and Lemeshow Test. Model 2(cutoff probability threshold 0.5), Model 3(cutoff probability threshold 0.4), and Model 5(PCA) were rejected because they showed high value of deviance(-2 Log likelihood) compared to Model 4.

Since Model 4 was best in giving the least deviance(-2 Log likelihood) and good PAC value, and interaction between variables was involved we have satisfied the assumption of non -linearity between independent variable and logit by checking the significance value.

There are other certain assumptions that cannot be ignored to accept Model 4 as the best model. The first assumption is the response variable should be 'Mutually exclusive'. In our dataset the response variable 'Pricecat' has two outcomes ie 'Expensive' or 'Budget' which does not seem to have had any commonalities between them. So, this assumption is not violated. The Next assumption is a check of 'Multicollinearity'. Independent variables should not be correlated with each other. This can be checked by VIF test. From Fig .34. all the predictor variables have value < 5 satisfying the assumption.

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	lotSize	.489	2.046
	landValue	.642	1.557
	livingArea	.250	4.002
	bedrooms	.477	2.094
	bathrooms	.478	2.091
	rooms	.398	2.511
	waterfront	.974	1.027
	lot_living_land	.383	2.609

a. Dependent Variable: PriceCat

Fig .34. VIF values of variables in Model 4

The third assumption is the Absence of Outliers/Influential data points this can be checked by the cook's distance. From Fig .35. it can be cross verified that no point is crossing above 0.2 all points are lying below 1. Hence no data points are influenced.

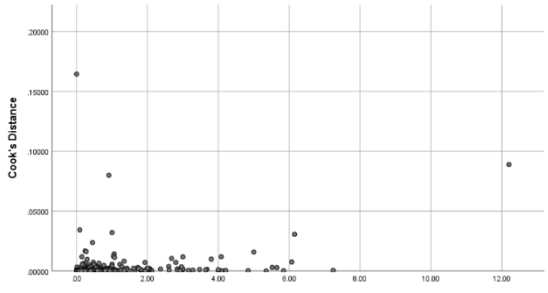

Fig .35. Cook's distance for Model 4

The last assumption is an adequate sample size. Logistics work well with a large number of samples minimum of 20 values per predictor is required. As our dataset holds 12 predictors minimum value required is 240, but we have 1709 values, this satisfies the sample size assumption.

Now let's interpret Model 4. From Fig .28. the Nagelkerke R Square also known as Pseudo R - Square value of 63.8% variation in criterion variable can be accounted to the predictor variables in the model. Fig .36. describes the contingency table of Model 4, this shows the values are almost equal for both the choices and the model adequately fits the data.

Contingency Table for Hosmer and Lemeshow Test

		PriceCat = Budget		PriceCat = Expensive		
		Observed	Expected	Observed	Expected	Total
Step 1	1	167	168.648	4	2.352	171
	2	169	163.103	2	7.897	171
	3	161	154.762	10	16.238	171
	4	137	139.472	34	31.528	171
	5	114	119.107	57	51.893	171
	6	87	91.763	84	79.237	171
	7	51	56.375	120	114.625	171
	8	34	28.770	137	142.230	171
	9	10	9.100	161	161.900	171
	10	2	.899	168	169.101	170

Fig .36. Contingency table of Model 4

Classification Table[a]

			Predicted		
			PriceCat		Percentage Correct
	Observed		Budget	Expensive	
Step 1	PriceCat	Budget	809	123	86.8
		Expensive	168	609	78.4
	Overall Percentage				83.0

a. The cut value is .500

Fig .37. Classification table of Model4

We can compare this Fig .37. classification table with Block 0 Fig .22. classification table to check improvement in the model by adding predictor variables. The Model Overall correctly classified cases are 83% which is greater than 54.5% from the Null model. The percentage in the first two rows provides information regarding Specificity 86.8% and Sensitivity 78.4%. The model exhibited good Specificity over Sensitivity.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	lotSize	.809	.170	22.610	1	.000	2.246	1.609	3.136
	landValue	.000	.000	132.143	1	.000	1.000	1.000	1.000
	livingArea	.003	.000	92.734	1	.000	1.003	1.002	1.003
	bedrooms			12.550	6	.051			
	bedrooms(1)	4.189	2.345	3.192	1	.074	65.986	.666	6535.701
	bedrooms(2)	1.490	2.036	.535	1	.464	4.435	.082	240.048
	bedrooms(3)	1.784	2.024	.778	1	.378	5.956	.113	314.313
	bedrooms(4)	1.529	2.017	.575	1	.448	4.613	.089	240.439
	bedrooms(5)	.723	2.037	.126	1	.723	2.061	.038	111.740
	bedrooms(6)	.322	2.402	.018	1	.894	1.379	.012	152.941
	bathrooms			60.814	8	.000			
	bathrooms(1)	-35.631	56830.073	.000	1	.999	.000	.000	.
	bathrooms(2)	-16.027	40176.888	.000	1	1.000	.000	.000	.
	bathrooms(3)	-15.556	40176.888	.000	1	1.000	.000	.000	.
	bathrooms(4)	-14.606	40176.888	.000	1	1.000	.000	.000	.
	bathrooms(5)	-14.535	40176.888	.000	1	1.000	.000	.000	.
	bathrooms(6)	-13.260	40176.888	.000	1	1.000	.000	.000	.
	bathrooms(7)	-13.035	40176.888	.000	1	1.000	.000	.000	.
	bathrooms(8)	1.513	41619.726	.000	1	1.000	4.540	.000	.
	rooms			11.370	10	.329			
	rooms(1)	-19.153	28275.013	.000	1	.999	.000	.000	.
	rooms(2)	-.523	.920	.323	1	.570	.593	.098	3.600
	rooms(3)	.666	.690	.931	1	.334	1.946	.503	7.527
	rooms(4)	.101	.672	.023	1	.881	1.106	.296	4.133
	rooms(5)	.554	.658	.709	1	.400	1.740	.479	6.324
	rooms(6)	.274	.645	.180	1	.671	1.315	.371	4.658
	rooms(7)	.694	.648	1.146	1	.284	2.002	.562	7.137
	rooms(8)	.230	.631	.133	1	.715	1.259	.366	4.334
	rooms(9)	.567	.637	.794	1	.373	1.764	.506	6.143
	rooms(10)	.376	.685	.300	1	.584	1.456	.380	5.577
	waterfront(1)	-3.087	1.118	7.621	1	.006	.046	.005	.408
	landValue by livingArea by lotSize	.000	.000	24.807	1	.000	1.000	1.000	1.000
	Constant	9.542	40176.889	.000	1	1.000	13929.029		

a. Variable(s) entered on step 1: lotSize, landValue, livingArea, bedrooms, bathrooms, rooms, waterfront, landValue * livingArea * lotSize .

Fig .38. Variables in Equation

This table Fig .38. shows which of the variables got a significant impact on our choice of house category. The first column is B(Beta) having coefficients value negative or positive for predictor variables and have a t-value and significance of t-value associated with each. Beta is also the predicted change in Log Odds- for 1 unit change in the predictor, there is Exp(B) change in the probability of the outcome As expected we don't see B values for categorical variables for which dummy columns were created.

The Exp(B) is the odds ratio of each column. If Odds Ration >1 then it says that the probability of falling into the house category as expensive group is greater than the probability of falling into the house category budget group ,if Odds Ratio <1 then it says that the probability of falling into the house category as expensive is lesser than the probability of falling into the house category budget group and if Odds Ratio = 1then it says that the probability of falling into house category as expensive is equal to the probability of falling into the house category budget group.

Interpreting the odds ratio for lotsize, we can say that the odds of a house category offereing Expensive for 1 unit change in lotsize are 2.246 times higher than those do not offer Budget for 1 unit change in lot size with a 95% CI of 1.609 to 3.136.

III. Conclusion And Future work

Time series analysis we had started exploring the pattern of time series, type of decomposition, and as our data had trend, season, and irregularity we had applied various models which suited our data like Seasonal Naïve, Holt Winters, ETS, and SARIMA. Holt Winter gave the least RMSE value hence it was selected as the best model for the accurate forecast of US e-commerce retail sales. For future enhancement, we can apply the damp parameter to Holt Winters and check accuracy.

Logistic regression for classifying house into expensive or budget, the Mode1 4 showed highest overall accuracy percentage of 83% and also this model showed lowest deviance(-2 log likelihood) of 1246.424. We have had also explored PCA to convert data to lower dimensions. For future enhancements, we can check with multiple probability cutoff values and a few non-significant dummy values can be removed. This can increase probability accuracy.

IV. References

[1] J. Brownlee, "How to Decompose Time Series Data into Trend and Seasonality," *Machine Learning Mastery*, Jan. 29, 2017. https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/ (accessed Jan. 01, 2022).

[2] S. Sirivella, "Simple Forecasting Methods," *Analytics Vidhya*, May 09, 2020. https://medium.com/analytics-vidhya/simple-forecasting-methods-a8016812ae38 (accessed Jan. 01, 2022).

[3] S. Date, "Holt-Winters Exponential Smoothing," *Medium*, Jul. 28, 2020. https://towardsdatascience.com/holt-winters-exponential-smoothing-d703072c0572 (accessed Jan. 01, 2022).

[4] A. Graves, "Time series forecasting with SARIMA model," *towardsdatascience.com*, Jan. 07, 2020. https://towardsdatascience.com/time-series-forecasting-with-a-sarima-model-db051b7ae459#:~:text=Per%20the%20formula%20SARIMA(p,lags%20of%20the%20stationarized%20series)&text=q%20and%20seasonal%20Q%3A%20indicate,seasonal%20length%20in%20the%20data (accessed Jan. 02, 2022).