Retail Sales Forecast and House Price Prediction Using Time Series and Logistic Regression

Kishore Lakshmanan
ID: 20253583
MSc in Data Analytics - B – 2021-2022
Statistics for Data Analytics
Terminal Assignment-Based
Assessment – Semester 1
National College of Ireland, IRELAND
Email:x20253583@student.ncirl.ie

***Abstract*— In the internet era, there is huge generation of large amount of data happening in every single day. So, there will be a lot of accumulation in the data which is left as unknown pattern. we cannot process this data without the help of some useful technologies. In this paper, we are going to look at the beneficial statistics concepts to derive prediction of future world from the sales and housing domain. We have implemented the various time series concept to check the better model among them for the Sales forecast. With the binomial logistic regression, we are going to predict the house price as its classified in terms of expensive or budget house.**

Keywords—Internet, generation, accumulation, pattern, technologies, statistics, prediction, time series, forecast, logistic regression.

I. Introduction

Sales Forecasting is the forecast of the future sales from the historical sales data. It is important to the company which is new arrival to the market or already existing one to experience the high growth in their sales by attracting their customers. So, they need to add lot of new products or new services to the market. It needs a supply capacity planning to overcome the unnecessary wastage of their material. In this aspect the sales forecast plays a major role in the industry.

There are several house crises going on all over the world. Most of the people are cheated by the property owners by selling their house with excessive rates. Hence the appropriate predictive model should involve in the housing sector to categories the houses based on their characteristic of the building to make decision whether it would be budget or expensive.

We are going to look at the various statistical techniques in below sections like description of data using descriptive statistics and visualization, model building process, diagnostic and assumption checking and summary of the model. For this project, we have used RStudio and SPSS software for the implementation of model.

II. Description of Data

A. Dataset Information

In this project, we are using 2 datasets to implement two different approaches. There is a e-commerce retail sales datasets which contains United States sales information which provides quarterly sales report from the year 1999, Q4 to the year 2021, Q2. This dataset is used to compare the 3 different time series model such as simple time series, exponential smoothing and ARIMA/SARIMA models.

Next dataset is the house categories which includes the data about the characteristics of houses sold in the United States during the certain period and based on the price scale, houses are classified into expensive or budget one. With this dataset we are going to implement the binary logistic regression with the best model to do our analysis.

B. Descriptive Statistics

RAW TIME SERIES DATA: There are two factors exists in this dataset which are date and ecomnsa. The data consists of quarterly data and ecomnsa column has the sales rate during that period. This dataset contains nearly 100 records from the previous sales. The below figure shows the raw time series data of ecommerce US retail sales data.

```
> class(ecomm_us)
[1] "data.frame"
> colnames(ecomm_us)
[1] "DATE"     "ECOMNSA"
> head(ecomm_us)
        DATE ECOMNSA
1 1999-10-01     5241
2 2000-01-01     5553
3 2000-04-01     6059
4 2000-07-01     6892
5 2000-10-01     9104
6 2001-01-01     7923
```

Fig 2.1. Raw time series data of E-commerce sales data

```
> ecomm <- ts(ecomm_us$ECOMNSA, frequency=4, start=c(1999,4), end=c(2021,2))
> class(ecomm)
[1] "ts"
> ecomm
      Qtr1   Qtr2   Qtr3   Qtr4
1999                      5241
2000   5553   6059   6892   9104
2001   7923   7816   7737  10784
2002   9621  10076  10760  14166
2003  12358  12973  13909  17915
2004  16201  16502  17371  22523
2005  20142  20953  22171  28121
2006  25490  25817  26892  35135
2007  30403  31589  32352  42126
2008  34270  34260  33486  39576
2009  32284  32924  34494  45805
2010  37059  38467  40075  54320
2011  44252  45442  46188  64504
2012  51822  52627  53878  73869
2013  58100  59753  60776  82929
2014  65242  68619  70071  94057
2015  74242  77934  79643 106418
2016  84079  88850  90483 120054
2017  95932 102797 104719 139773
2018 111673 119007 120311 155124
2019 124384 134714 140538 178865
2020 141521 193624 191573 235957
2021 196808 211704
```

Fig 2.2. Transformation of data frame to time series data

The raw time series data then transformed to time series data to forecast the sales using various time series model. The above shown figure is the transformed data from the data frame class to the time series class. It has been created by the ecomnsa variable, setting frequency as 4 which means a quarterly data, start and end data of the time series. This gives us the overall time series data of the ecommerce sales starting from the year Q4, 1999.

HOUSE CATEGORIES: In this dataset, there are two types of variable available such as continuous and categorical type. These variables are also known as scale level data and nominal data respectively. The descriptive statistics is used to describe the variables in the dataset using the mathematical notations such as mean, median and other such factors which are shown below. The following figure shows the descriptive statistics of both continuous and categorical variables separately.

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
lotSize	1709	.00	12.20	.4932	.01615	.66772	.446	7.140	.059	80.729	.118
age	1709	0	225	27.76	.699	28.895	834.949	2.522	.059	7.659	.118
landValue	1709	200	412600	34758.35	849.853	35132.968	1234325407	3.092	.059	16.096	.118
livingArea	1709	616	5228	1756.88	14.981	619.308	383542.843	.908	.059	1.298	.118
pctCollege	1709	20	82	55.67	.249	10.289	105.868	-1.054	.059	.651	.118
bathrooms	1709	.0	4.5	1.905	.0159	.6583	.433	.311	.059	-.438	.118
Valid N (listwise)	1709										

Fig 2.3. Descriptive statistics of continuous variables

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
bedrooms * PriceCat	1709	100.0%	0	0.0%	1709	100.0%
fireplaces * PriceCat	1709	100.0%	0	0.0%	1709	100.0%
rooms * PriceCat	1709	100.0%	0	0.0%	1709	100.0%
fuel * PriceCat	1709	100.0%	0	0.0%	1709	100.0%
waterfront * PriceCat	1709	100.0%	0	0.0%	1709	100.0%
newConstruction * PriceCat	1709	100.0%	0	0.0%	1709	100.0%

bedrooms * PriceCat

Crosstab

Count

		PriceCat		
		Budget	Expensive	Total
bedrooms	1	4	3	7
	2	299	45	344
	3	485	328	813
	4	125	357	482
	5	16	37	53
	6	2	6	8
	7	1	1	2
Total		932	777	1709

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Interval by Interval	Pearson's R	.409	.021	18.509	.000[c]
Ordinal by Ordinal	Spearman Correlation	.431	.020	19.735	.000[c]
N of Valid Cases		1709			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Fig 2.4. Descriptive statistics of categorical variable

In statistics, we use descriptive method to illustrate the continuous variables and cross tab or frequency is used to analyze the categorical variable. The above figure shows the relation between the categorical variable of bedroom and price variable. It also shows the missingness value of the variables in the dataset.

C. Visualization of data

Visualization technique is the best method to elaborate the different type of variable in the graphical manner to easy understanding of the users. In this section, we are going to look at the data visualization of two different dataset used in this project.

TIME SERIES: Using R, we are going to visualize the time series data with inbuild packages such as ggplot which is available in the fpp2 library. In this we are using various plots and methods to visualize the data. The below graph shows the time series data starting from the year 2010 to visualize the graph in detail. For this we used the subset method with the window function which starts the graph with the provided input as a start date.

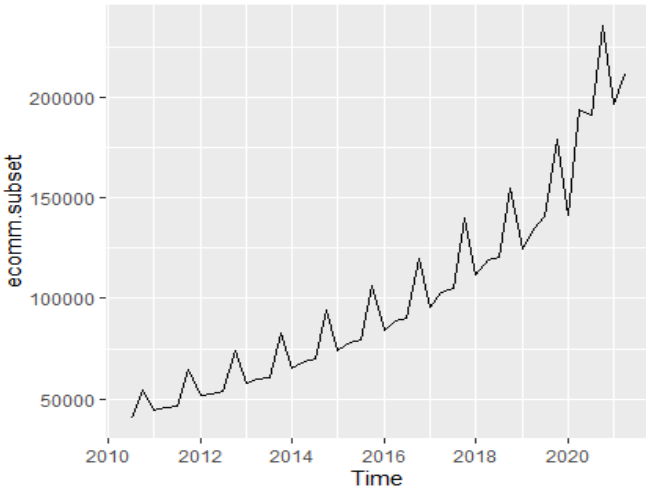


Fig 2.5. Time Series Data Using Subset Method

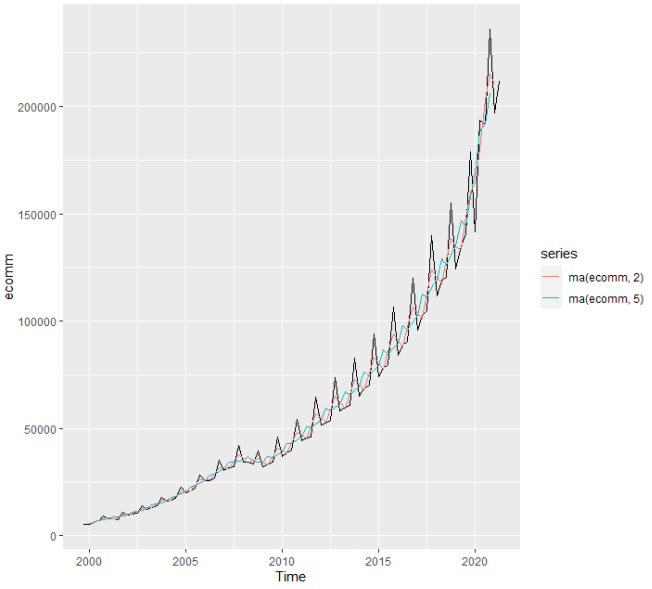


Fig 2.6. Simple Moving Average of Time Series Data

The subset graph doesn't show the exact ups and downs in the sales data. So, we are planning to the simple moving method to show the accuracy of the plot. For that we need to adjust the ma based on the accurate plot we received. In this dataset, we get the ma as 2 which is quite accurate compared to the all-other ma. We need to overcome the underfit when it is less smoothened and overfit as its more smoothened.

HOSING CATEGORIES: Using SPSS, we are going to look at the basic visualization techniques to understand the data.

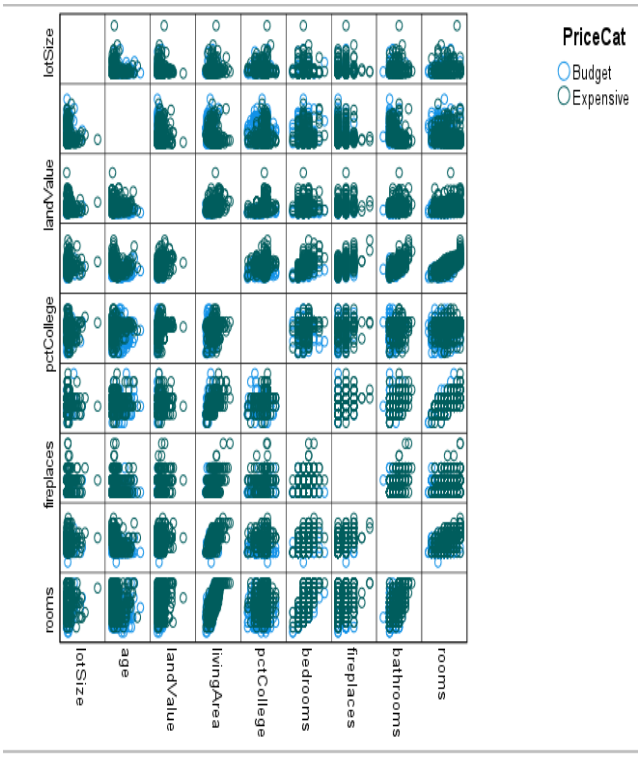


Fig 2.7. Overview of housing data using Scatterplot

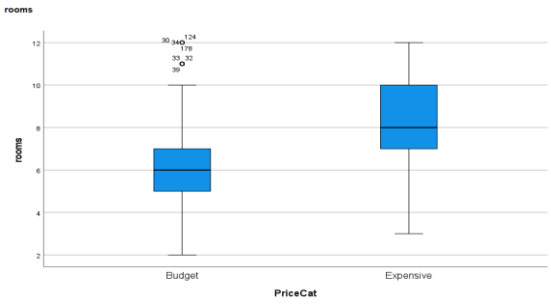


Fig 2.8. Box Plot diagram for room vs price category

The Box plot diagram shows that the rooms that are normally distributed to the price category based upon the whisker's extension to the both sides of boxes. This plot can be useful for visualizing large amount of data.

The histogram would determine the data whether the variables are normally distributed or not. This may be skewed sometimes. The below graph shows the relation between room and price category. In this both of them are normally distributed in nature. Hence, we can assume there is a strong relationship between these two variables.

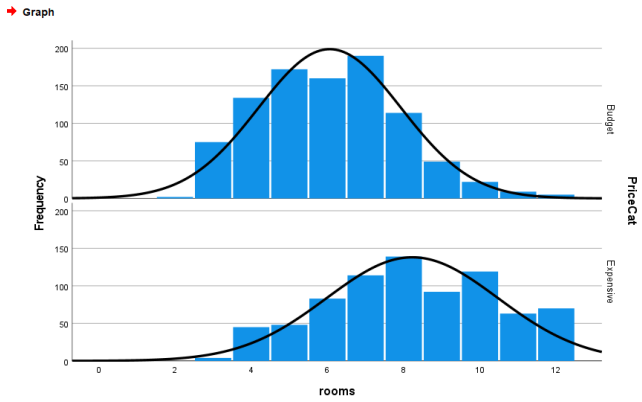


Fig 2.4. Histogram chart for room vs price category

D. Use cases

In this section, we are going to see the research questions made available from these 2 datasets.

1. Predict the US E-Commerce sales forecast for three period ahead.
2. Predict the US House price prediction whether it is budget or expensive based on the house characteristics.

E. Objectiveness

The main aim of this project is to analyze the various time series methods and best fit model for the binary logistic regression. With these two techniques, we are going to predict the sales movement in future 3 periods and to predict the house prices using different house characteristic in the United States Region. The performance measure of all these model has to be analyzed to get the good model.

III. RELATED WORK

There are lots of studies taken prior in the field of sales and housing sectors using various machine learning algorithms and statistics. In this section, we are going to see the literature review of some of the papers.

A. Jain [1] uses data mining techniques to forecast sales and prediction. They estimate that the extreme gradient boosting algorithm is useful for making prediction model accurately. Also, they compared the XGBoost predictor with the traditional regression algorithms. Armstrong J.S [2] was examined alternative strategies and casual approaches which are recommended for making decisions in forecasting. Jiang [3] uses the feasibility of time series model, hybrid model and machine learning model for predicting Walmart sales. They conclude by stating that machine learning model performs well in the sales prediction. G. Nunnari [4] uses the residual time series by nonlinear auto regression methods by using neuro-fuzzy and feed-forward neural networks method. Akshay [5] used regression techniques to forecast the sales and found that the boosting algorithm have better results.

Raga [6] uses the different regression techniques to predict the house prices and found out the best among them. Dueganjali [7] has implemented the house price prediction using various classification algorithms and with that models they compared to get the best model. Manasa [10] uses the

linear regression and boosting algorithm to check the best performance model by comparative analysis between the model and this is used to predict the price based on the factors that affects the price. Wang [11] was used deep learning model for house price prediction using heterogeneous data with the help of joint self-attention mechanism. Dhillon [12] has analyzed the Airbnb prices using machine learning techniques and they concluded that the random forest method was the best one out of all other techniques.

IV. MODEL BUILDING PROCESS

In this part, we have to build our model which should be best for finding our research questions. The following content shows the what are all steps we have taken to pick the final model, intermediate model, model transformation for correcting any issues in the model and final model. We have to implement two types of techniques in two datasets. Firstly, we see the model building process of time series method and followed by the binary logistic regression.

TIME SERIES MODEL

A. Steps taken before applying the time series model

Before applying any of the model, we need to analyze the graph to check the fluctuation, trend and seasonality of the time series data. It has some irregular or error components in the data. So, removing this fluctuation we need to plot the graph using moving average to see the clear view of the plot. It has been done and shown in the descriptive statistics part.

Time series data have a seasonal aspect that can be decomposed to trend, seasonal and irregular component. In seasonal components, there are two types additive and multiplicative. Majority will use the second option.

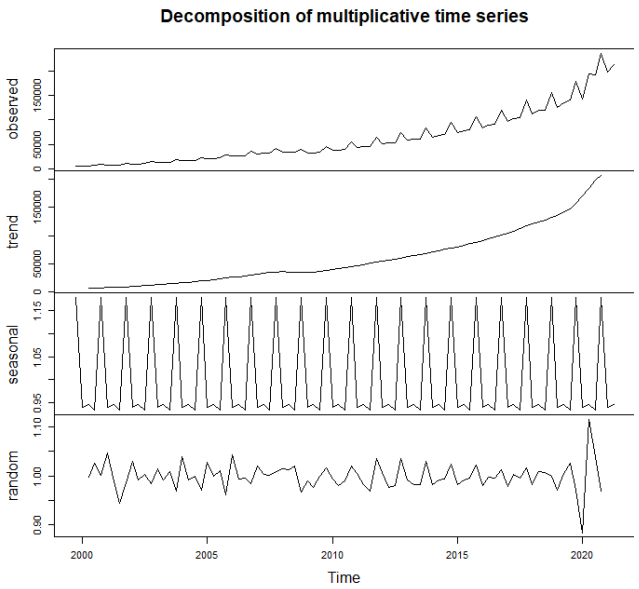


Fig 4.11 Decomposition of multiplicative time series

Here we use the Loess STL decomposition method to see the trend, seasonal and remainder in single plot and overcomes the drawback of classical model.

```
> exp(fit.stl$time.series)
         seasonal      trend remainder
1999 Q4 1.1757525   4757.462 0.9369641
2000 Q1 0.9449193   5467.295 1.0748811
2000 Q2 0.9569574   6260.874 1.0112846
2000 Q3 0.9405822   7097.404 1.0324024
2000 Q4 1.1757525   7677.328 1.0085705
2001 Q1 0.9449193   8035.552 1.0434682
2001 Q2 0.9569574   8289.479 0.9852914
2001 Q3 0.9405822   8641.946 0.9518407
2001 Q4 1.1757525   9181.772 0.9989356
2002 Q1 0.9449193   9901.996 1.0282596
```

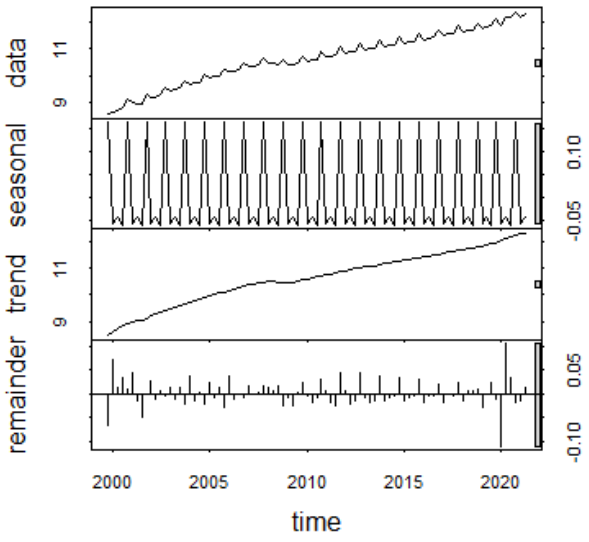


Fig 4.12 Decomposition using STL

B. Simple time series model

Simple time series model is used to forecast the time series in exactly simple ways. There are 3 types of models in it which includes average(mean) method, naïve method and seasonal naïve model.

```
attr(,"class")
[1] "meanf"

Error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE      ACF1
Training set  0 54131.92 43030.09 -144.599 175.031 4.334126 0.9100178

Forecasts:
        Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
2021 Q3          62673 -8045.15 133391.2 -46181.78 171527.8
2021 Q4          62673 -8045.15 133391.2 -46181.78 171527.8
2022 Q1          62673 -8045.15 133391.2 -46181.78 171527.8
```

Fig 4.21 Mean model of simple time series

The above model shows the 3 period ahead forecast by mean model and accuracy of this model also explained.

```
Model Information:
Call: naive(y = ecomm, h = 3)

Residual sd: 15390.3391

Error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 2400.733 15390.34 9385.384 2.855386 13.10789 0.9453254 -0.5980073

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3         211704 191980.5 231427.5 181539.5 241868.5
2021 Q4         211704 183810.7 239597.3 169044.9 254363.1
2022 Q1         211704 177541.9 245866.1 159457.5 263950.5
```

Fig 4.22 Naïve model of simple time series

The above model shows the three period ahead forecast by the naïve model with its accuracy of this model.

```
Model Information:
Call: snaive(y = ecomm, 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
```

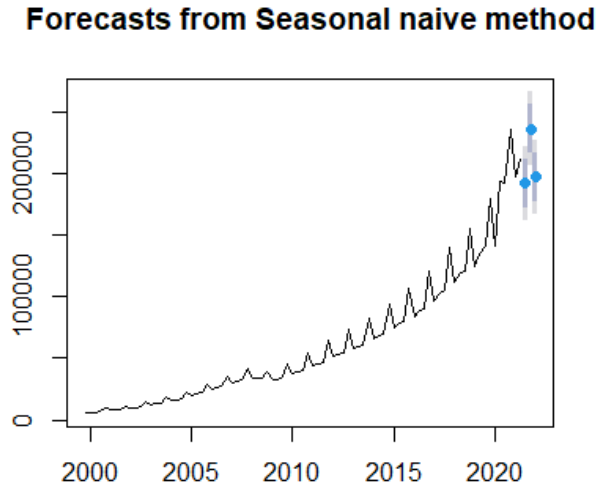


Fig 4.23 Seasonal Naïve model

Seasonal naïve model is shown above with its accuracy rate of the model. Out of 3 simple time series model, seasonal time series model works better. So, we can say there is some sort of seasonal trend involved in the time series model. We can drill further by other models to find the best fit.

C. Exponential Smoothing model.

Exponential smoothing model widely used for the short-term predictions in wide range of applications. It gives higher weight for most recent observation. This is dived into 3 types. They are single exponential smoothing, holt exponential smoothing and holt-winters exponential smoothing.

Simple exponential smoothing fits a time series model with levels and irregular components. The below graph shows its 3 period ahead forecast and accuracy of the model.

```
> ecomm_ses
        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

Simple exponential smoothing

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

  Smoothing parameters:
    alpha = 0.5447

  Initial states:
    l = 6946.7401

  sigma:  13348.08

     AIC     AICc      BIC
2045.359 2045.648 2052.757
> round(accuracy(ecomm_ses),2)
                ME      RMSE      MAE  MPE  MAPE MASE  ACF1
Training set 4244.2 13193.76 7275.81 5.42 10.34 0.73 -0.28
```

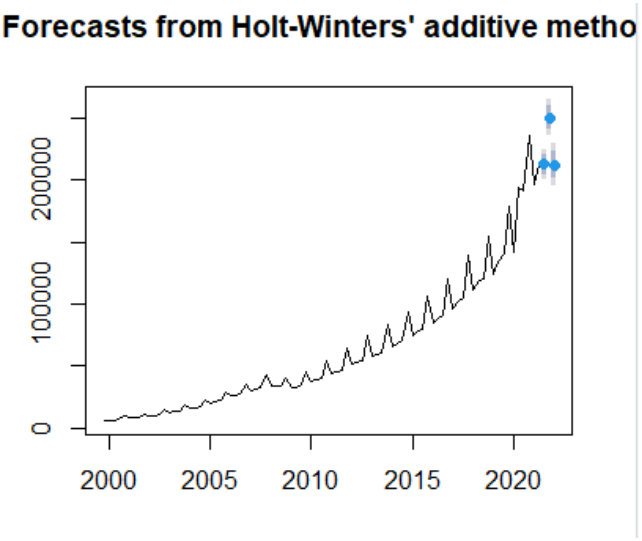
Holt exponential model fits a time series with the level and trend. The below model describes the 3 period ahead and accuracy of the model.

```
> ecomm_holt
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3         227347.1 212818.8 241875.3 205128.0 249566.1
2021 Q4         239504.2 224640.6 254367.8 216772.3 262236.2
2022 Q1         251661.4 236190.1 267132.7 228000.1 275322.7
> ecomm_holt$model
Holt's method

Call:
 holt(y = ecomm, h = 3)

  Smoothing parameters:
    alpha = 0.1367
    beta  = 0.0794

  Initial states:
    l = 4806.0515
    b = 700.8892

  sigma:  11336.45

     AIC     AICc      BIC
2018.865 2019.605 2031.194
> round(accuracy(ecomm_holt),2)
                ME      RMSE      MAE MPE MAPE MASE  ACF1
Training set 1658.16 11072.77 6649.48 0.2 9.72 0.67 -0.15
```

It also fits by using the ets method (A,A,N) model which implements the additive trend component in the model.

Holt-winter exponential smoothing model which fits the model with level, trend and seasonal components. The below figure shows its 3-period forecast and its accuracy.

```
     AIC     AICc      BIC
1916.303 1918.641 1938.496
> ecomm_hw
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3         212619.4 204731.1 220507.7 200555.3 224683.5
2021 Q4         250146.0 240662.6 259629.3 235642.4 264649.5
2022 Q1         212203.7 201050.2 223357.3 195145.9 229261.6
> accuracy(ecomm_hw)
                   ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set 830.2474 5865.446 2924.569 1.008276 10.86375 0.2945718 -0.02026113
```



D. ARIMA model

Arima model aims to states the autocorrelation of the autocorrelations in the data using the lags(autoregressive), lags of forecast errors (moving average) and difference in the time series(integrated). The covariance stationary is the term in which mean, variance do not change over the time. The difference is 1 and p and q can be found out by the acf and pacf method. With the several test we have the component for fit model for arima.

```
> accuracy(fit)
                   ME     RMSE      MAE      MPE     MAPE      MASE      ACF1
Training set 4992.366 11140.06 6917.789 7.686931 10.54427 0.7370811 0.1994871
> #Forecasting with the fitted model
> forecast(fit, 3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3         189787.7 175428.3 204147.0 167827.0 211748.4
2021 Q4         217086.2 201073.6 233098.9 192597.0 241575.5
2022 Q1         200010.2 182022.4 217998.0 172500.2 227520.2
```

The model has the p, d, q of (3, 1, 0) which is best fit of ARIMA model by evaluating the various tests.

E. SARIMA

SARIMA is the seasonal arima which gives the forecast of the seasonal trend by using the auto.arima function. This gives the best method for the seasonal arima. The below shown figure is the seasonal arima with best P,D,Q value with its accuracy.

```
> ecommfit<-Arima(ecomm,order = c(1,1,0),seasonal = c(1,1,0))
> ecommfit %>%forecast(h=3) %>% autoplot()
> forecast(ecommfit,h=3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       218050.7 210942.3 225159.2 207179.3 228922.2
2021 Q4       257560.8 248937.1 266184.5 244372.0 270749.6
2022 Q1       219880.7 209609.5 230152.0 204172.2 235589.3
> accuracy(ecommfit)
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 784.1104 5318.922 2390.89 0.2915821 3.651811 0.2408179 0.006567584
```

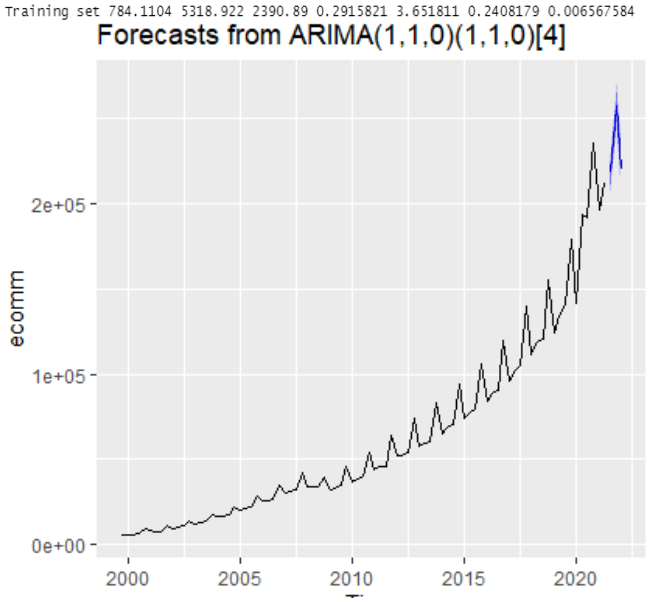


Fig 4.24 FORECAST PLOT USING SARIMA

LOGISTIC REGRESSION MODEL

A. Steps taken for achieving final regression model

This is the first step has to be taken before applying any of the model. There are many steps which we consider before applying model that are variable selection,

1) Variable transformation and Selection:

In variable transformation, there are three factors that are exist as a string type as yes / no and other types. Fuel, waterfront and construction variables which are in string type which has been transformed to the nominal data using the automatic recode function available in the SPSS software. The below shown figure is describe how the transformation takes place in the three variables.

Next step is to check the correlation between the variables to eliminate the highly correlated one out from our model to get rid of the multicollinearity issues. In this dataset, we have some combinations which is correlated more than 0.5 which are living area & bathroom (.721 correlation), living area & room (.733 correlation), living area & bedroom (.656 correlation). Living area is more correlated with all other factors stated above. The value which is near to 1 are highly correlated. So we need to eliminate such factors from our model.

```
AUTORECODE VARIABLES=newConstruction
  /INTO Newconstructionl
  /BLANK=MISSING
  /PRINT.
newConstruction into Newconstructionl
Old Value   New Value   Value Label

No                  1   No
Yes                 2   Yes


AUTORECODE VARIABLES=waterfront
  /INTO Waterfrontl
  /BLANK=MISSING
  /PRINT.
waterfront into Waterfrontl
Old Value   New Value   Value Label

No                  1   No
Yes                 2   Yes


AUTORECODE VARIABLES=fuel
  /INTO fuell
  /BLANK=MISSING
  /PRINT.
fuel into fuell
Old Value   New Value   Value Label

electric            1   electric
gas                 2   gas
oil                 3   oil
```

Fig 3.11 Transformation of variables

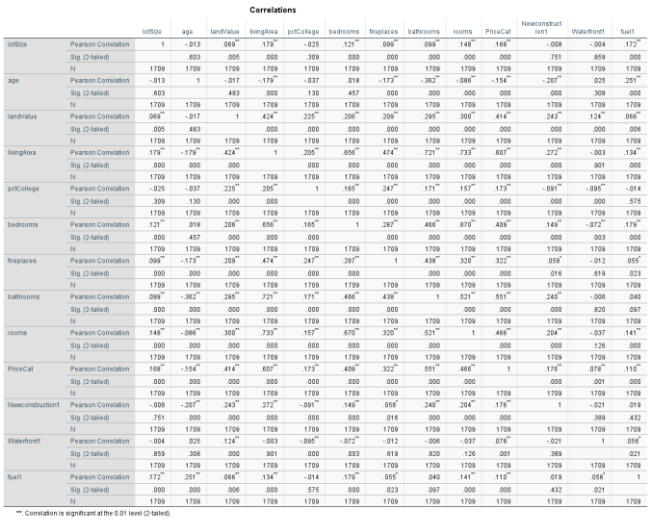
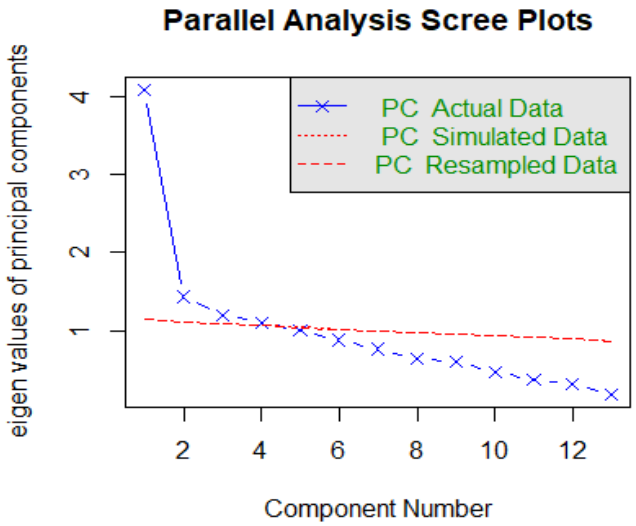


Fig 3.12. Correlation between the dependent and independent variables

2) Dimension Reduction Techniques

In this dataset, we have more than 10 factors, so it is difficult to implement our model with all the combination. The main think is to avoid correlation factors in our model. To avoid this issue, we are going into the dimension reduction method to overcome the more pairwise correlation. The 10 correlated factors are divided into 4 uncorrelated components using pca and factor method.

The below scree plot shows that the which component having the more eigen value. Based upon that we select the number of components should present in our model. These factors are then rotated to extract the components includes the factors. The following graph shows the component analysis of three components.



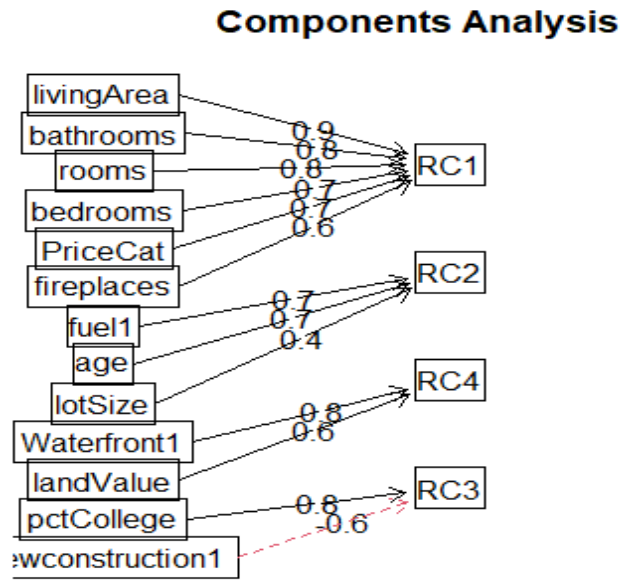


Fig 3.13 Components extraction using dimension reduction

3) Model Building

In the model building process, we are starting the model with the simple model and then we can proceed with the future model by analysing the diagnostic steps. We are going to check the accuracy of the model with and without adding the PCA features. The below simple model build using the simple factors.

Sample model has the PseudoR square vale of 58% and some variables are highly correlated. We can see from the p value of the factors in the model below.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1377.640[a]	.436	.582

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

```
Call:
glm(formula = PriceCat ~ landValue + livingArea + pctCollege +
    Newconstruction1 + fuel1, family = binomial, data = house)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0780  -0.5951  -0.2586   0.5711   2.7859

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -6.347e+00  6.763e-01  -9.385   <2e-16 ***
landValue        4.016e-05  3.633e-06  11.057   <2e-16 ***
livingArea       3.187e-03  1.771e-04  17.994   <2e-16 ***
pctCollege      -1.488e-02  7.386e-03  -2.015   0.0439 *
Newconstruction1  1.675e-01  4.376e-01   0.383   0.7018
fuel1             3.565e-02  1.276e-01   0.279   0.7800
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2355.1  on 1708  degrees of freedom
Residual deviance: 1377.6  on 1703  degrees of freedom
AIC: 1389.6

Number of Fisher Scoring iterations: 6
```

Classification Table[a]

Observed			Predicted PriceCat 0	Predicted PriceCat Budget	Percentage Correct
Step 1	PriceCat	0	799	133	85.7
		Budget	184	593	76.3
	Overall Percentage				81.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	landValue	.000	.000	122.256	1	.000	1.000
	livingArea	.003	.000	323.777	1	.000	1.003
	pctCollege	-.015	.007	4.062	1	.044	.985
	Newconstruction1	.168	.438	.147	1	.702	1.182
	fuel1	.036	.128	.078	1	.780	1.036
	Constant	-6.347	.676	88.070	1	.000	.002

a. Variable(s) entered on step 1: landValue, livingArea, pctCollege, Newconstruction1, fuel1.

Fig 3.14 Sample model building process

B. Intermediate Models

There are many intermediate models which we have build in this stage to predict the best fit of the model. Some of the sample models are stated below.

```
house1<-glm(PriceCat~landValue+room+pctCollege+Newconstruction1+fuel1,data = house, family=binomial)
summary(house1)
```

Above model also having the multicollinearity issue and the accuracy of the model also very less for the prediction.

```
house7<-glm(PriceCat~bedroom+room+pctCollege+Newconstruction1+fuel1,data = house, family=binomial)
summary(house7)

house8<-glm(PriceCat~bedroom+landvalue+bathroom+room+pctCollege+Newconstruction1+fuel1,data = house, fami
summary(house8)
```

Some of the intermediate models are shown below which has the same above stated problems. All the model has the different type of issues which needs to be sort out to get the perfect model.

C. Treatment of multicollinearity and transformation undertaken

All of our model has some issues which are normality, multicollinearity etc. For these we have undertaken the steps which is dimension reduction. These will convert the correlated variables into smaller uncorrelated components. By building our model through the factor and PCA will removes the multicollinearity issues.

Rotated Component Matrix[a]

	Component			
	1	2	3	4
lotSize				
age		.698		
landValue			.591	
livingArea	.908			
pctCollege				.756
bedrooms	.734			
fireplaces	.582			
bathrooms	.806			
rooms	.796			
PriceCat	.698			
Newconstruction1				-.618
Waterfront1			.814	
fuel1		.730		

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

a. Rotation converged in 7 iterations.

Below model is our final binary logistic model which have been built through these four components and passes almost all the tests. This model gives the Pseudo R square value of 84%, PAC value of 94%, sensitivity and specificity both are having 94%.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1682.248	4	.000
	Block	1682.248	4	.000
	Model	1682.248	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	672.852[a]	.626	.837

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	298.136	8	.000

Classification Table[a]

			Predicted		
			PriceCat		Percentage Correct
	Observed		0	Budget	
Step 1	PriceCat	0	876	56	94.0
		Budget	47	730	94.0
	Overall Percentage				94.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 4 for analysis 1	-1.093	.137	63.819	1	.000	.335
	REGR factor score 3 for analysis 1	5.087	.330	237.714	1	.000	161.828
	REGR factor score 2 for analysis 1	.056	.109	.258	1	.612	1.057
	REGR factor score 1 for analysis 1	4.237	.220	369.805	1	.000	69.205
	Constant	.597	.116	26.583	1	.000	1.816

a. Variable(s) entered on step 1: REGR factor score 4 for analysis 1, REGR factor score 3 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1.

Fig 3.31 Final Logistic Regression Model

V. DIAGNOSTICS AND ASSUMPTION CHECKING

TIME SERIES MODEL

A. Augmented Dickey-Fuller Test

This test is used to evaluate the assumption of stationary in our time series data. The alternative function for this test is ndiffs function. The p value of less than .r is good model. Our model has the p value of .1. So it is passed in this test.

```
        Augmented Dickey-Fuller Test

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

Fig 5.1 Augmented Dickey-Fuller Test

B. Normal Distribution

The graph shown below represents the normality of the model. Our model is considered as quite normal not that perfect. Since we have selected the SARIMA model based on the auto ARIMA model which gives this result.

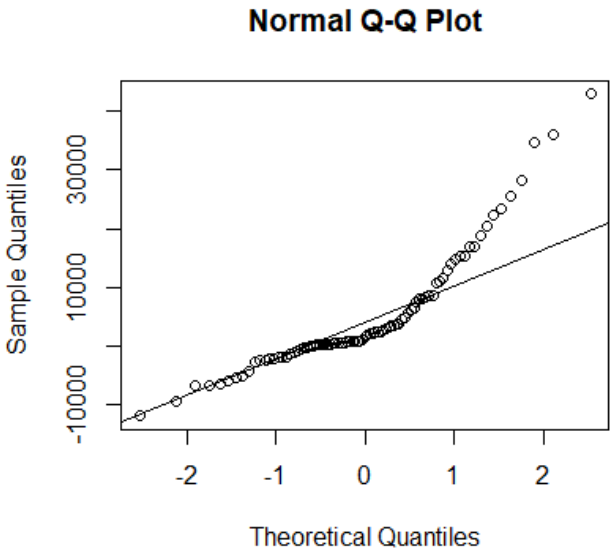


Fig 5.2 Normal Distribution

C. Ljung Box Test

This test evaluates the residuals in the model. Below shown is the value of residuals and its plot. By seeing the p value and the graph it passed this test too.

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

        Box-Ljung test

data:  ecommfit$residuals
X-squared = 0.0038835, df = 1, p-value = 0.9503

> checkresiduals(ecommfit)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,0)(1,1,0)[4]
Q* = 3.375, df = 6, p-value = 0.7605

Model df: 2.   Total lags used: 8
```

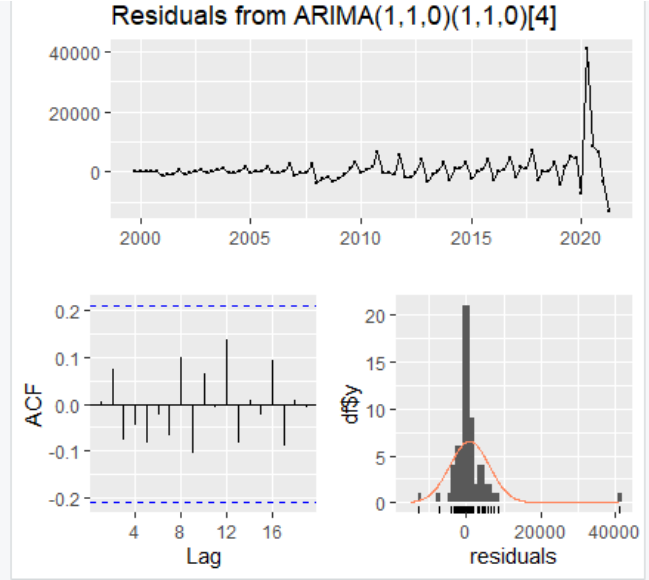


Fig 5.3 Residuals Plot of SARIMA model

LOGISTIC REGRESSION MODEL

D. Omnibus Test

This test is used to check that the null hypothesis of coefficient of all the variables are zero. If all the variables are zero then its distributed x square with 1 degrees of freedom.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1682.248	4	.000
	Block	1682.248	4	.000
	Model	1682.248	4	.000

Fig 5.4 Omnibus test

E. Hosmer & Lemshow Test

This test is useful to check how much best the model is fit. The value less than .05 is indicates the poor fit. The below shown is the best fit of our model.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.054	8	.261

Fig 5.5 Hosmer & Lemshow Test

VI. MODEL SUMMARY AND VISUALIZATION

In this section, we are going to look at the performance fit of our model and visualization of use cases provided as aim of the project.

First of all, we look at the time series model and its overall performance of the model. The below shown descriptive are the accuracy measures of the time series model.

```
[1] "meanf"

Error measures:
                    ME     RMSE      MAE      MPE     MAPE     MASE      ACF1
Training set  0 54131.92 43030.09 -144.599 175.031 4.334126 0.9100178
```

```
Call: naive(y = ecomm, h = 3)

Residual sd: 15390.3391

Error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 2400.733 15390.34 9385.384 2.855386 13.10789 0.9453254 -0.5980073
```

```
Call: snaive(y = ecomm, 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
```

```
> round(accuracy(ecomm_ses),2)
                   ME     RMSE      MAE  MPE  MAPE MASE  ACF1
Training set 4244.2 13193.76 7275.81 5.42 10.34 0.73 -0.28
> round(accuracy(ecomm_holt),2)
                   ME     RMSE      MAE MPE MAPE MASE  ACF1
Training set 1658.16 11072.77 6649.48 0.2 9.72 0.67 -0.15
> accuracy(ecomm_hw)
                    ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set 830.2474 5865.446 2924.569 1.008276 10.86375 0.2945718 -0.02026113
> accuracy(ecomm_ets)
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 938.2113 5397.091 2039.609 1.088049 3.051819 0.2054358 0.04771821
> accuracy(fit)
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 4992.366 11140.06 6917.789 7.686931 10.54427 0.7370811 0.1994871
> accuracy(ecommfit)
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 784.1104 5318.922 2390.89 0.2915821 3.651811 0.2408179 0.006567584
```

Fig 6.1 Overall Performance measures of the Time Series

From the above representation, it shows the SARIMA performs well as 5318 RSME values and second is auto ets function has 5397 RSME. Both has a least difference in the performance measure. Last is the simple mean model has 54131 RSME value. Since the time series data is a seasonal one. This can be predicted beforehand by seeing the below ggplots.

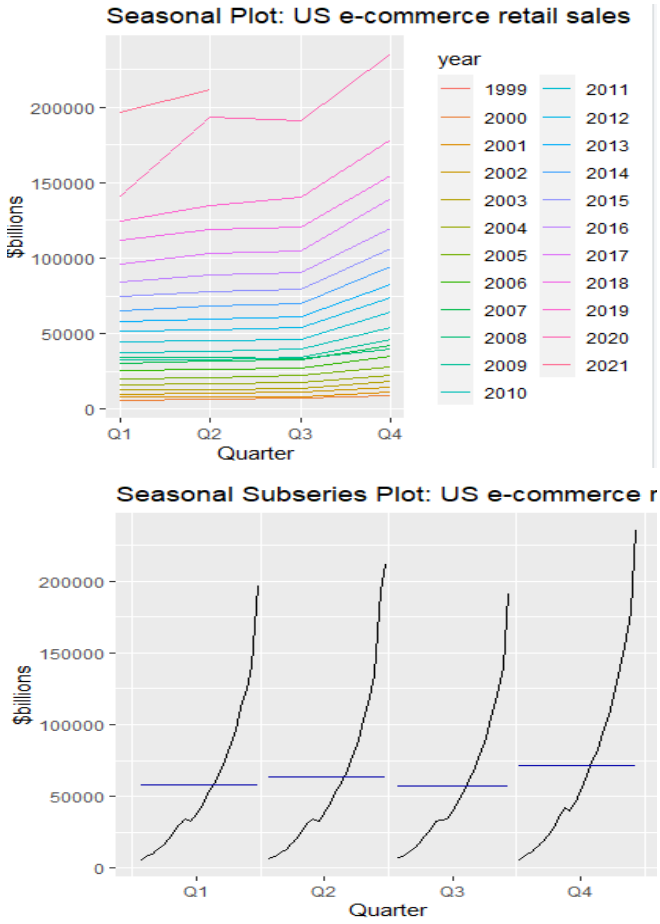


Fig 6.2 Visualization of best quarter using ggseasonalplot

The above graph shows the best quarter using ggseasonalplot which shows the 4th quarter has the most sales happening period. The following is the 3 period ahead forecast of the best fit model.

```
> ecommfit %>%forecast(h=3) %>% autoplot()
> forecast(ecommfit,h=3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 9
2021 Q3       218050.7 210942.3 225159.2 207179.3 228922.
2021 Q4       257560.8 248937.1 266184.5 244372.0 270749.
2022 Q1       219880.7 209609.5 230152.0 204172.2 235589.
```

Fig 6.3 Forecast of sales using best fit time series model (SARIMA)

In the binary logistic regression part, we could see that the performance is more when using the dimension reduction techniques. Below is our best fit model with factor as a four component from PCA. Here Pseudo R square value is considered as accuracy. This model has the accuracy of 84%.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	672.852[a]	.626	.837

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001

By the classification table, we see the specificity, sensitivity and PAC value which is 94% for all. By the odds ratio in the variables, we could see the component 3 has the highest odds ratio. In this prediction, variables in the components 3 are the major factors of the prediction. 1 unit change in the component 3, there will be a 161 times greater the house prices.

Classification Table[a]

Observed			Predicted PriceCat 0	Predicted PriceCat Budget	Percentage Correct
Step 1	PriceCat	0	876	56	94.0
		Budget	47	730	94.0
	Overall Percentage				94.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 4 for analysis 1	-1.093	.137	63.819	1	.000	.335
	REGR factor score 3 for analysis 1	5.087	.330	237.714	1	.000	161.828
	REGR factor score 2 for analysis 1	.056	.109	.258	1	.612	1.057
	REGR factor score 1 for analysis 1	4.237	.220	369.805	1	.000	69.205
	Constant	.597	.116	26.583	1	.000	1.816

Fig 6.4 Best fit of Logistic regression model

VII. CONCLUSION

Our project focuses on model building and diagnostic measures taken for the time series and logistic regression techniques to find out the good model. We have achieved this part and provides reasonable summary and visualization to showcase our use cases of our project with interpretation of the best model. In the sales forecast prediction, SARIMA model performs well compared to other time series model. Since, the time series data was moving into the seasonal trend all over the periods. In the house price prediction, we have gone through the several model and analyzed its performance. But the model with PCA of dimensionality reduction makes the model perfect and the accuracy was achieved higher compared to all other intermediate model.

REFERENCES

[1] A. Jain, M. N. Menon, and S. Chandra, "Sales Forecasting for Retail Chains," *www.semanticscholar.org*, 2015.

[2] J. Armstrong, "Sales Forecasting. The IEBM Encyclopedia of Marketing," 1999.

[3] H. Jiang, J. Ruan, and J. Sun, "Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast," IEEE Xplore, Mar. 01, 2021.

[4] G. Nunnari and V. Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study," *2017 IEEE 19th Conference on Business Informatics (CBI)*, Jul. 2017, doi: 10.1109/cbi.2017.57.

[5] A. Krishna, A. V, A. Aich, and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Dec. 2018, doi: 10.1109/csitss.2018.8768765.

[6] CH. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," *IEEE Xplore*, Mar. 01, 2019.

[7] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," *IEEE Xplore*, Mar. 01, 2019.

[8] https://uc-r.github.io/descriptives_categorical

[9] https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b

[10] J. Manasa, R. Gupta, and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," *IEEE Xplore*, Mar. 01, 2020.

[11] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism," *IEEE Access*, vol. 9, pp. 55244–55259, 2021, doi: 10.1109/ACCESS.2021.3071306.

[12] J. Dhillon *et al.*, "Analysis of Airbnb Prices using Machine Learning Techniques," *IEEE Xplore*, Jan. 01, 2021.