Application of Time Series and Logistic Regression Models through Statistical Approach

Pramod Ramu
StudentId: 20205759
Statistics For Data Analytics, MSc in Data Analytics
National College of Ireland Dublin, IRELAND
Email: x20205759@student.ncirl.ie. URL: www.ncirl.ie

***Abstract*—This study encompasses quarterly time series of United States e-commerce retail sales which commences from quarter 4 and house categories data set which include various characteristics of the house sold in US region.So for the given time series simple time series model,exponential time series model and ARIMA/SARIMA models were applied and the optimal model was choosed on the lower RMSE values.On the other hand for house categories data set logistic regression model was applied and based on the house characteristics the house was classified as expensive or budget and the accuracy was calculated using confusion matrix.**

***Index Terms*—Time Series , Simple Time Series Model , Exponential Smoothing Model , ARIMA , SARIMA , RMSE , Logistic Regression , Confusion Matrix.**

I. Introduction

The time series is a sequential set of data points that occur repeatedly in some period of time.A time series is used to forecast an event that have been chosen at a certain period of time where the data points have been recorded at regular interval of time.The time series could be applied on any variables that change on time [1].So this can be used for analysis as well as forecasting.Both the analysis and forecasting can be performed by obtaining the historical values or patterns [2].So this project encompasses quarterly time series of United States e-commerce retail sales which commences from quarter 4. So the objective is to apply Simple time series models,exponential time series models and Auto Regression Integrated Moving Average(ARIMA) as well as SARIMA models for estimating the optimal model for the series.

A logistic regression model is usually applied on dependent variable which is dichotomous variable.For example whether the customer is satisfied or not, whether the team will win the match or not,so basically there should be two choice in the variable.So a logistic regression model classifies based the dependent variable based on the probability of the event that happen with one or more independent variables.It measures the importance of the independent variables on influencing the dependent variable.This project includes the data set of house categories.It provides the details of the characteristics of the houses houses sold in a region of USA at particular period.So the objective is to build a logistic regression model which help to study the characteristics of the house so that the house would be classified as 'expensive' or 'budget'.

The methodology section encompasses analysis of the data provided,detailed information on building the models,model evaluation as well as various diagnostic tests.The conclusion parts gives gives an optimal model for the given time series as well as overall summary of all the implemented models in both time series and logistic regression.

II. Methodology

The process of analysis of the given data,building and evaluation of the model were carried out using R [3].

A. Analysis of the given time series

The given data was the quarterly time series of the United States e-commerce in billions.The series has two components date and sales values.The series starts from fourth quarter of the year 1999 and ends at second quarter of the year 2021.In order to start the process the given series was placed in a time series object using ts() function and plot was created using autoplot function figure1.By look at the plot it seems that the series is a combination of both season and trend as we can observe there is fall near the year 2010.The season and season subseries season plots of the time series were also created in the figure2 and in the figure3.To remove the irregular and error component from the time series simple moving average method was used .The figure4 shows smoothing of the time series using simple averages method. In order to examine further the series was decomposed using seasonal decomposition methods to check the different patterns of the data.By looking the plot of the series it seems that seasonal decomposition using multiplicative model suits better for for this series.Initially the model was decomposed using classic approach of seasonal decomposition using multiplicative model but it was found some missing values in first and last rows of the series.So seasonal and trend decomposition using Loess method was used using stl() function.So for multiplicative models log transformation of the time series object must be taken.The series after taking log transformation is shown in the figure5.So the series was decomposed using season and trend decomposition method.This plot in figure6 shows the original,seasonal,trend and other patterns of the time series.The next step was to apply simple time series model.

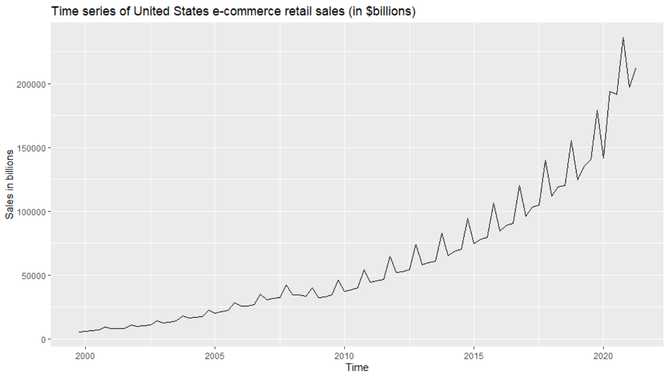


Fig. 1. Plot of the given time series

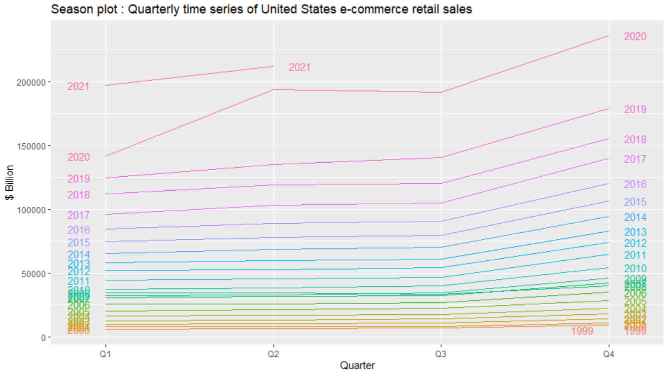


Fig. 2. Season plot of the given time series

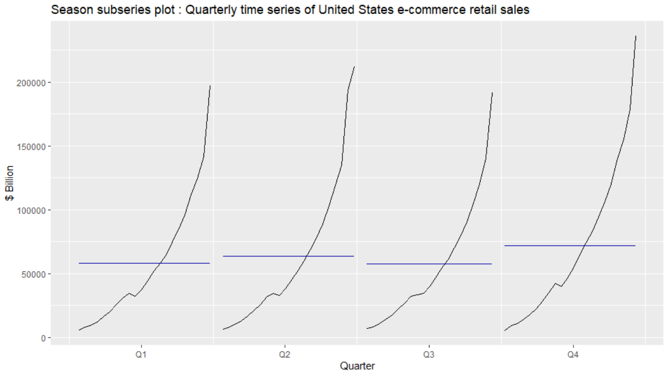


Fig. 3. Season subseries plot of the given time series

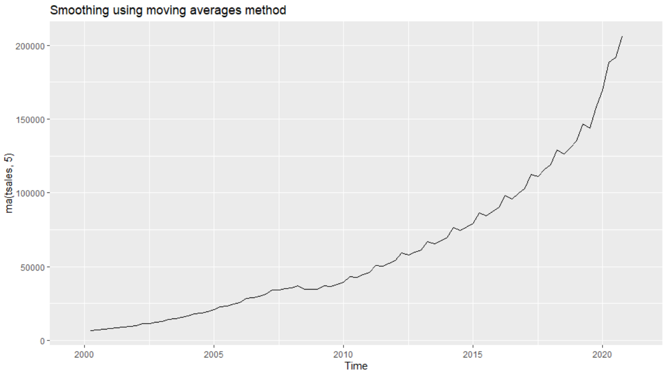


Fig. 4. Smoothing using simple moving averages method

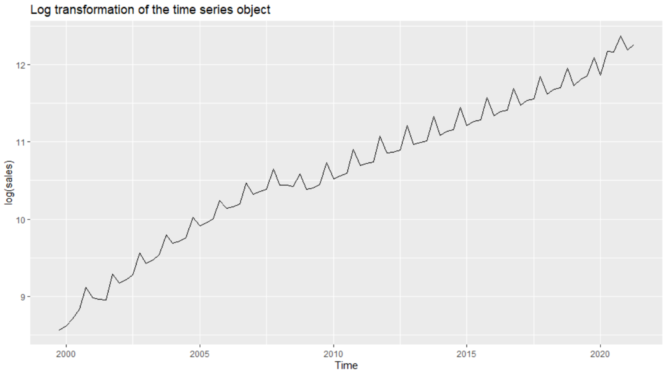


Fig. 5. Time series plot after taking log transformation

B. Building and evaluation of simple time series model

In order to forecast the future sales simple time series models were applied to this time series.The models which were used were mean model,naive model and seasonal naive model.In the mean model the forecasts of the future values are computed based on the mean or average of the historical data.In naive model the forecasts of the future values are the values of the last observation.This model is also called as random walk model.In the seasonal naive model the forecasts of the future values are equal the values of the last observation from the same season of the year.The accuracy is computed for the training set.The summary of the mean model is shown in the figure7.The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 0 , 54131.92 , 43030.09 , -144.599 , 175.031 , 4.334126 and 0.9100178 respectively.For the accuracy purpose importance is given to RMSE value.The summary also gives the forecasts of the future values for three periods.The plot of the model is shown in the figure8.

The summary of the naive model is shown in the figure9.The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 2400.733 , 15390.34 , 9385.384 , 2.855386 , 13.10789 , 0.9453254 and -0.5980073 respectively.The summary also gives the forecasts of the future values for three periods.The plot of the model is shown in the figure10.

The summary of the seasonal naive model is shown in the figure11.The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 9786.711 ,15144 , 9928.205 , 15.44986 , 15.85113 , 1 and 0.8350162 respectively.The summary also gives the forecasts of the future values for three periods.The plot of the model is shown in the figure12.

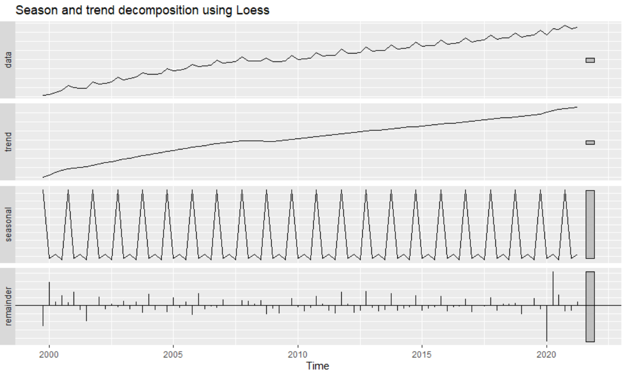
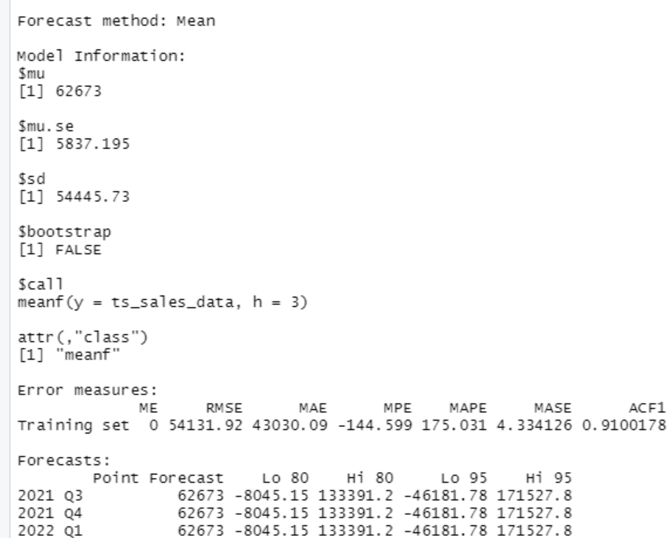


Fig. 6. Season and trend decomposition using Loess



```
Forecast method: Mean

Model Information:
$mu
[1] 62673

$mu.se
[1] 5837.195

$sd
[1] 54445.73

$bootstrap
[1] FALSE

$call
meanf(y = ts_sales_data, h = 3)

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. 7. Summary of the mean model

C. Building and evaluation of exponential smoothing models

Another approach for forecasting sales was applying the exponential smoothing models.The different exponential models were simple exponential smoothing(ses) model,Holt linear trend method and Holt-Winters exponential smoothing model.The simple exponential smoothing has only the the 'level'(alpha) component which is the smoothing parameter.It fits the series at level and forecasts are calculated based on weighted averages.The Holt linear trend method encompasses level(alpha) and trend(beta) components.It fits the time series at both level and trend.Holt-Winters exponential smoothing model has level(alpha),trend(beta) and seasonal(gamma) components.It fits the series at level,trend and seasonal.So here simple exponential smoothing model,Holt exponential smoothing model and Holt-Winters exponential smoothing model were implemented.For verifying the best method ets() function was used with 'ZZZ' as the model parameter.The accuracy is computed for the training set.

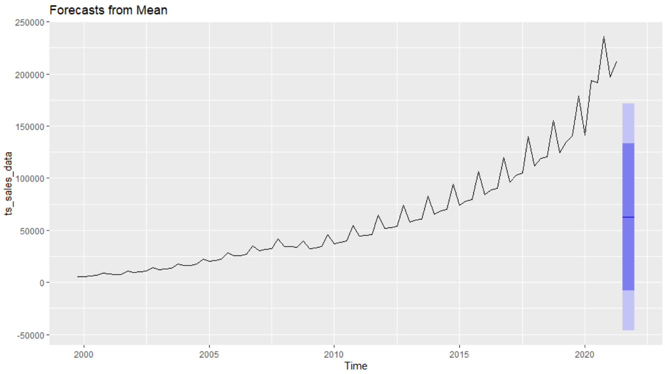
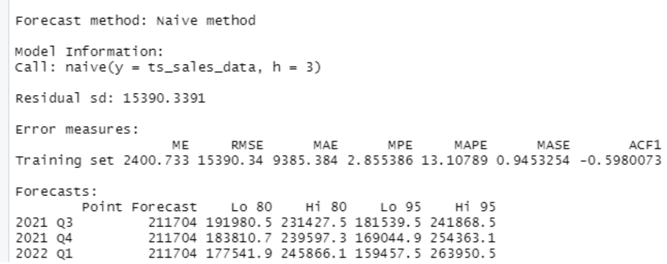


Fig. 8. plot of the mean model



```
Forecast method: Naive method

Model Information:
Call: naive(y = ts_sales_data, 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. 9. Summary of the naive model

The summary of the ses model is shown in the figure13.From the summary it seems that the alpha value is 0.5447 .The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 4244.196 , 13193.76 , 7275.815 , 5.423953 , 10.33686 , 0.7328429 and -0.284854 respectively.The AIC value observed was 2045.359.The summary also gives the forecasts of the future values for three periods.The plot of the model is shown in the figure14. The summary of the Holt exponential smoothing model is shown in the figure15.From the summary it seems that the alpha and beta values were 0.1367 and 0.0794 .The AIC value observed was 2018.865.The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 1658.159 , 11072.77 , 6649.48 , 0.1999953 , 9.721311 , 0.6697565 and -0.1500883.The summary also gives the forecasts of the future values for three periods.The plot of the model is shown in the figure16. The next model which was applied was Holt winters exponential smoothing model.So initially two models were built using both additional and multiplication seasonality components and the seasonality and the best method applicable to this series was verified by using ets() function.From the summary of the Holt Winters model of additive method in figure20 it can inferred that the alpha, beta and gamma values were 0.5904 ,0.0769 and 0.4096 .The AIC value of the model was 1916.303.The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 830.2474 , 5865.446 , 2924.569 , 1.008276 , 10.86375 , 0.2945718 and -0.02026113.On the other hand from the summary of the Holt Winters model of multiplicative method in the figure figure21.The alpha, beta and gamma values were 0.4666 ,

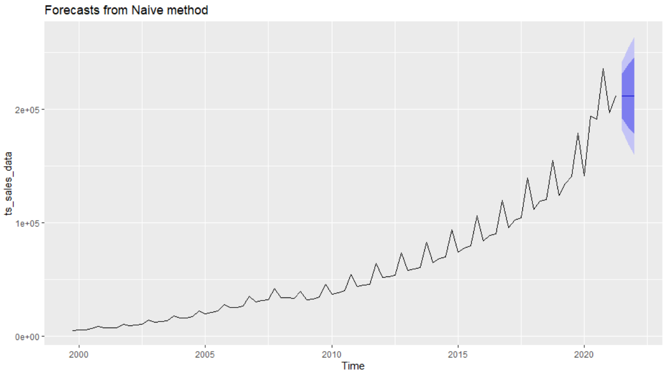
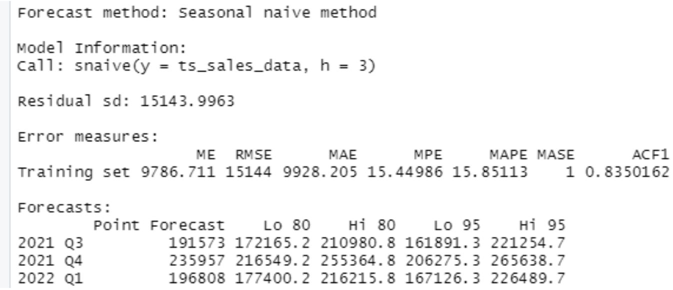


Fig. 10. plot of the naive model



```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = ts_sales_data, 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. 11. Summary of the seasonal naive model

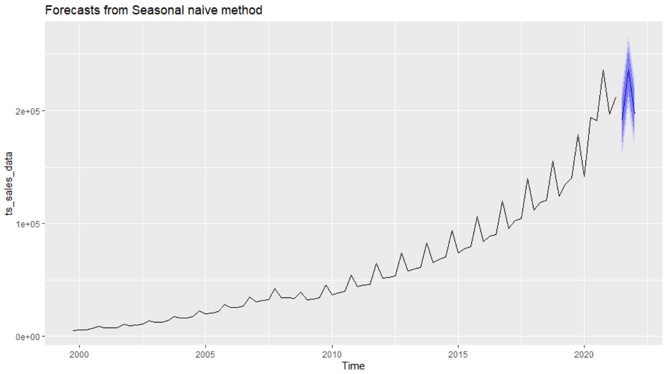
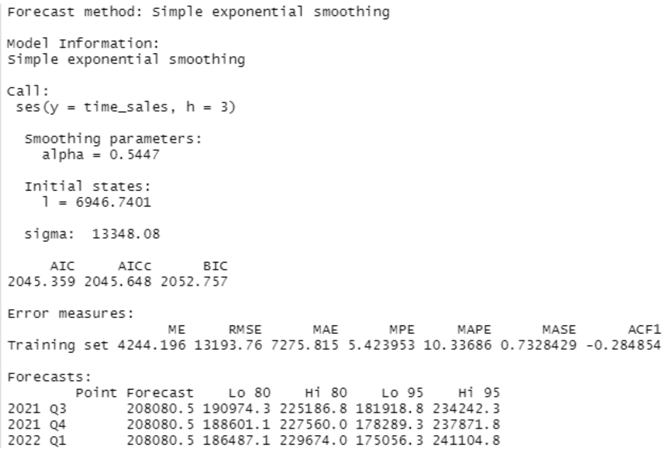


Fig. 12. plot of the seasonal naive model



```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = time_sales, 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. 13. Summary of simple exponential smoothing model

0.1178 and 1e-04.The observed AIC value was 1768.792.The ME , RMSE , MAE , MAPE , MASE and ACF1 values of the model were 925.311 , 4969.59 , 2321.772 , 0.8432752 , 4.250947 , 0.2338562 and 0.182943.The summary also encompasses the forecasts of the future values for three periods. The plot of Holt Winter exponential smoothing model encompasses both additive and multiplicative forecast as shown in the figure22. Along with fitting the time series with these models ets() function with model parameter as 'ZZZ' used to check the best method applicable for this series.The ets function() showed the 'MAM' model which is the multiplicative Holt Winters method with multiplicative level,additive trend and multiplicative seasonal component suits best for the given time series.The summary of the model using ets() function is shown in the figure20.From the summary it can inferred that the alpha, beta and gamma values were 0.6684,0.0577 and 0.2913 .The AIC value of the model was 1726.852.The ME , RMSE , MAE , MAPE , MASE and ACF1 values were 938.2113 , 5397.091 , 2039.609 , 1.088049 , 3.051819 , 0.2054358 and 0.04771821.The summary also gives the forecasts of the future values for three periods.The plot of the model is shown in the figure21.

D. Building and evaluation ARIMA and SARIMA models

ARIMA stands for Auto Regressive Integrated Moving Average and it would be specified using the parameters p,d and q.Where p is auto regression value AR(p),d is number of differencing required I(d) and q is the moving average MA(q).These are the non-seasonal components.On the the other hand SARIMA encompasses seasonal components P,D and Q along with non-seasonal components. For the given time series both ARIMA and SARIMA models were fitted.The first model implemented was the ARIMA.The given model had both trend and seasonality components.In order to make it stationary,the components were differenced and implemented Augmented Dickey - Fuller Test(adf test).If the P-value observed in adf test was statistically significant then it can be said that the the series is staionary.So the given series was checked with number of differences required by using 'ndiff'.It showed 1 and the series was differenced using 'diff' function and at the same the adf test was also implemented.It was observed the p-value was not statistically significant and so in order to resolve this second order differencing was taken.Now on implementing the adf test it was observed the the p-value was statistically significant.So it was confirmed that the series is stationary.So to get the p and q values pacf(partially auto-correlated function) and acf(auto-correlated function) functions were implemented.It was a bit tricky in identifying the p and q values,hence auto-arima function was used to fit only the non-seasonal

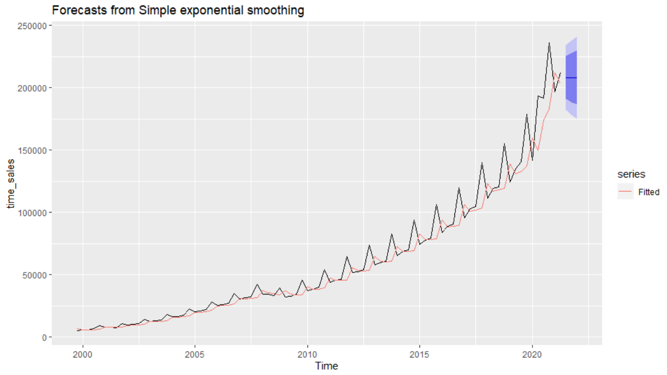
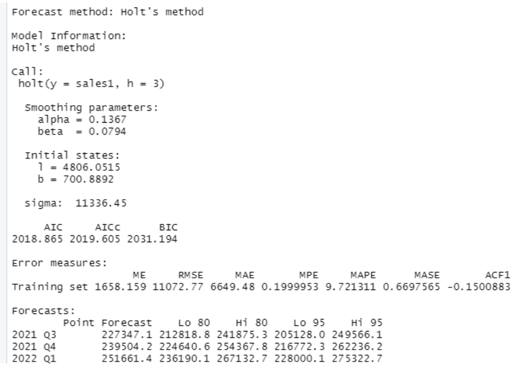


Fig. 14. Plot of ses model



```
Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = sales1, 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

Error measures:
                   ME     RMSE     MAE       MPE     MAPE      MASE       ACF1
Training set 1658.159 11072.77 6649.48 0.1999953 9.721311 0.6697565 -0.1500883

Forecasts:
        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
```

Fig. 15. Summary of Holt exponential smoothing model

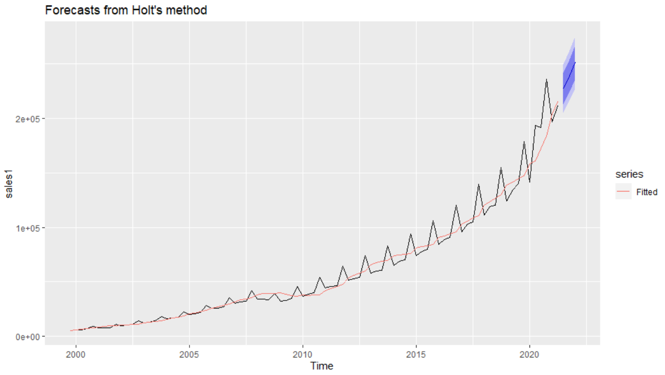
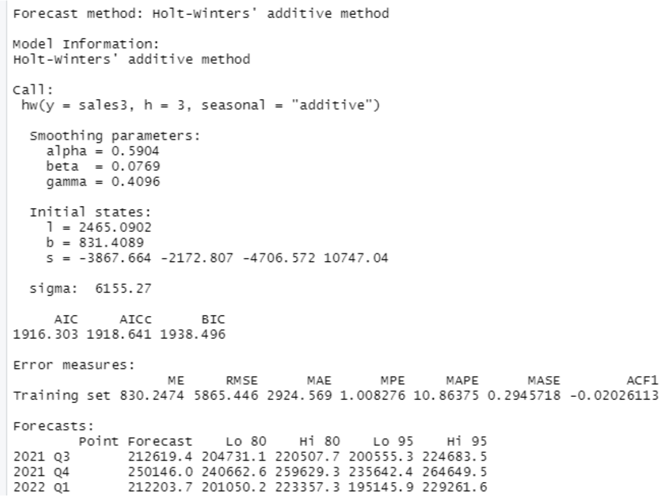


Fig. 16. Plot of Holt exponential smoothing model



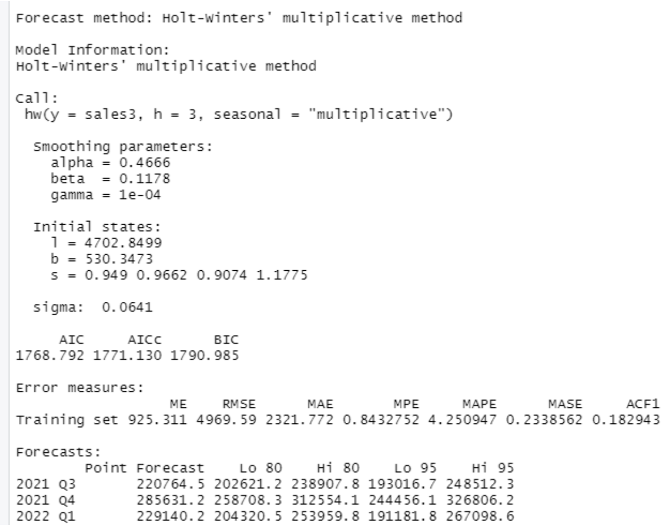
```
Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

Call:
 hw(y = sales3, h = 3, 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

Error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set 830.2474 5865.446 2924.569 1.008276 10.86375 0.2945718 -0.02026113

Forecasts:
        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
```

Fig. 17. Summary of Holt Winter additive method

components.The value of p,d and q was 1,1 and 0.It was found that the residuals were not normally distributed.Also box test was implemented using Ljung-box method,on checking the residuals the p-value was not statistically significant.The ME,RMSE,MAE,MPE,MAPE,MASE and ACF1 values were 3606.231,12639.1,7750.747,5.046993,11.06812,0.7806796 and -0.07246849.The AIC value was 1873.89 and the value of the auto regressive component(ar1) was -0.5603. So now the SARIMA model was applied to the series by incorporating the seasonal components.The values of the seasonal components P,D and Q was indentified by the auto-arima function and the values were 1,1 and 0.So now the SARIMA model was applied on the series with both non seasonal and seasonal components (p,d,q) and (P,D,Q).It was also found by implementing qqnorm plot that the residuals were normally distributed and also by implmenting the box test using Ljung-box method figure23 the result had a non-statistically significant values for the residuals in figure25 and in figure26.So it was confirmed that the SARIMA models fits best for the given time series as it encompasses seasonal component.From the model summary figure22. it was inferred that the auto regressiove (ar1) and seasonal auto regressive component values(sar1) were -0.3132 and -0.6250.The ME,RMSE,MAE,MPE,MAPE,MASE and ACF1 values were 784.1104 , 5318.922 , 2390.89 , 0.2915821 , 3.651811 , 0.2408179 and 0.006567584 respectively.The AIC value of the model was 1652.59.Also the plot for the SARIMA model and forecasts for 3 periods is shown in the figure27 and figure28

E. Analysis of the US house categories data

The given house categories data set consists of 1709 observations and 13 columns.The overall idea is to using the different characteristics of the house and need to classify the house as 'expensive' or 'budget',so the price-category will be the dependent variable which belongs to the dichotomous category.The characteristics of the house are lot size, age, land value, living area, percentage of local residents with college education,number of bedrooms, total number of fireplaces, number of bathrooms, number of rooms, fuel used for heating system, waterfront property and whether there is a new construction or not.The categorical variables observed were fuel,waterfront and new construction.The first step was to check the correlation between all the continuous variables using Pearson method in figure29.Also various plots for correlation was created in figure30 and figure31 .So by ob-



```
Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = sales3, h = 3, 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

Error measures:
                   ME    RMSE      MAE       MPE     MAPE      MASE     ACF1
Training set 925.311 4969.59 2321.772 0.8432752 4.250947 0.2338562 0.182943

Forecasts:
        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. 18. Summary of Holt Winter multiplicative method

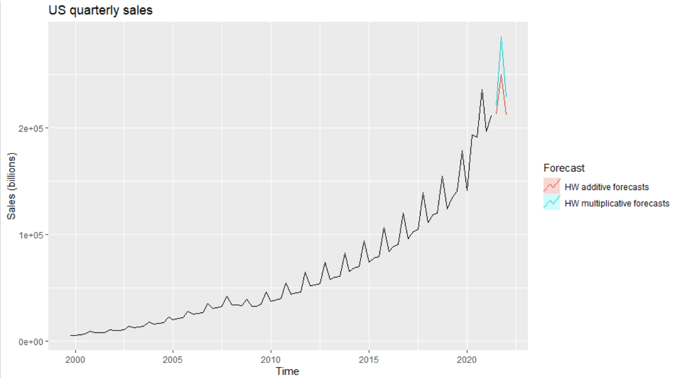
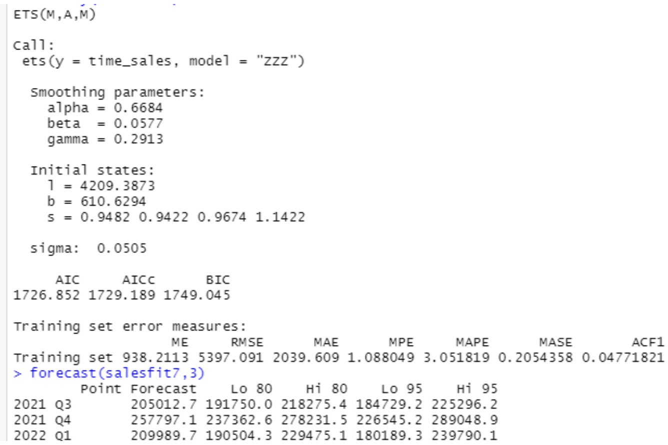


Fig. 19. Plot Holt winter exponential smoothing model

serving the correlation values none of the variables have bigger correlation value.So the next step was to check the descriptive statistics of the given data.The descriptive statistics is shown in the figure32.Different metrics such as mean,standard deviation,median,min,max,range,skew and kurtosis can be inferred from the descriptive statistics.By observing the descriptive statistics it seems the variables 'landvalue' has the highest and the variable 'lotsize' has the mean,median and the standard deviation.Also the variable 'livingarea' have a bigger values of mean,median and standard deviation.The higher skew and kurtosis values was observed in 'waterfront' variable where as 'pctcollege' has the lowest skew value and lowest kurtosis value was observed in price-category variable.The dependent variable 'pricecat' was in numerical type and it was converted to factor type and it also satisfies one of the assumption of the logistic regression that the dependent variable must be categorical variable.



```
ETS(M,A,M)

Call:
 ets(y = time_sales, 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

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 938.2113 5397.091 2039.609 1.088049 3.051819 0.2054358 0.04771821
> forecast(salesfit7,3)
        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. 20. Summary of the model using ETS function

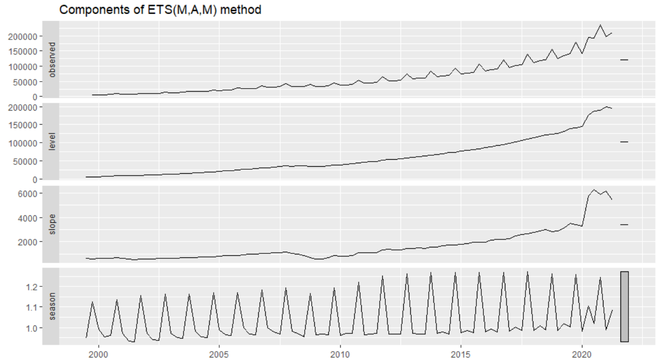
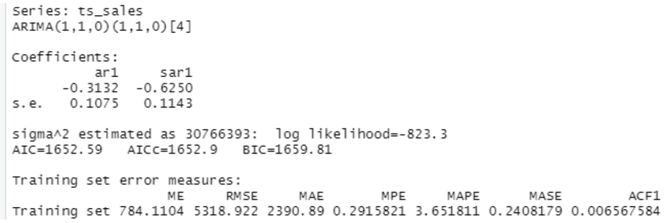


Fig. 21. Plot of the model which was build using ETS

F. Building Logistic Regression Model

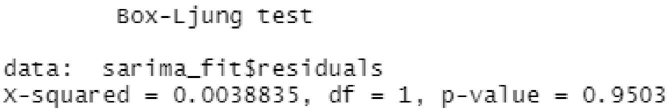
The logistic regression model was built using glm() function.For building the logistic regression model two approaches were followed.One by regular approach by including all the independent variable and removing the variables one by one which were not statistically significant.The second approach was by using dimensionality reduction.In this approach the independent variables were grouped into different components and using those components for building the model.

In the first approach,first all the independent variables were included in the glm() function with 'priceCat' as the dependent variable.It was found that the variable 'fuel' had the highest p-value,so it was dropped.Now the second model was build using all the independent variables but excluding 'fuel'.Now it was found that the 'rooms' variable had the highest p-value.So it was dropped and the third model was built excluding 'rooms' variable.Now the 'fireplaces' variable had the highest p-value,so it was dropped.The same process was followed again by building fourth,fifth,sixth and seventh models and the variables dropped in order were 'newConstruction','bedrooms','pctCollege' and 'age'.So the independent variables of the final model were 'lotSize','landValue','livingArea' and 'waterfront'.All these



```
Series: ts_sales
ARIMA(1,1,0)(1,1,0)[4]

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

sigma^2 estimated as 30766393:  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 784.1104 5318.922 2390.89 0.2915821 3.651811 0.2408179 0.006567584
```

Fig. 22. Summary of SARIMA model



```
         Box-Ljung test

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

Fig. 23. Ljung test for SARIMA model

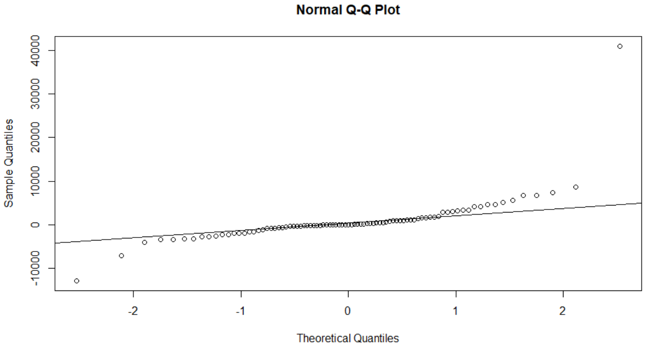
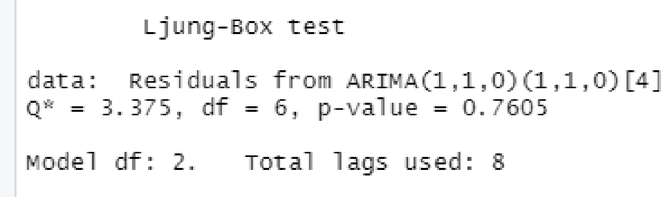


Fig. 24. Normal distribution for residuals using qqplot



```
        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. 25. Residual summary for SARIMA model

variables were statistically significant. In the second approach dimensionality reduction concepts were used to build the model.Here the significant continuous independent variables were grouped using principle component analysis approach.Before starting the principal component analysis the KMO and barlet test was performed for all the continuous variables in the data set.The overal measure of sampling adequacy was 0.8.The summary of KMO test is shown in the figure figure33.Along with this the Bartlett test was also conducted.The summary of the Bartlett test is shown in the figure34.The p-value was statistically significant.The scree plot is shown in the figure35.The analysis of the PCA is shown in the figure36 and figure37. Here the variables ’livingarea’ , ’rooms’ , ’bedrooms’ , ’bathrooms’ and ’fireplaces’ were grouped into first component ’rc1’,the ’age’ variable is the second component and the variables ’pctcollege’ , ’lotSize’ and ’landValue’ belong to the third component.Now the logistic regression model was built along with the these three groups and the categorical variables ’fuel’,’waterfront’ and ’newconstruction’.Initially it was found that the ’fuel’ variable had the highest p-value.So it was removed from the model.Now the new model was built using the three groups and with categorical variables ’waterfront’ and ’newconstruction’.Now it was observed that the ’newconstruction’ variable had the highest p-value.Hence it was removed from the model.The new model was built using the three groups RC1,RC2 and RC3 as well as the ’waterfront’ variable.Now it was found that all the variables were statistically significant and the AIC observed was 1501. Now the Hosmer lemeshow test was conducted for both the models.It was found that the for the first model the p-value was less than 0.05 which indicated the poor fit and for the second model the p-value was 0.6945 in the figure38 which supported the good fit.So the model which was built using PCA was chosen as the final model.The summary of the final model is shown in the figure39. The goodness of fit was also identified by pseudo rsquare values of various metrics like McFadden,CoxSnell,Nagekkerke with the help of PseudoR2 method.The summary is shown in the figure40. The next step was to satisfy the assumptions of the final logistic regression model.The dependent variable was made as factor variable of two choices ’budget’ and ’expensive’ with levels ’1’ and ’2’.This satisfied the assumption of the dependent variable should be dichotomous variable.As the total number of observations was 1709 and this satisfied the assumption of that there should be large samples of the data.Next assumption was there should be no multicollinearity.This assumption was tested by using vif function.It was observed that for the variables the vif values were not greater than 2.It is shown in the figure41.This shows that there is no multicollinearity.Another assumption was to verify that there should be no influential data points.This was tested by plotting influential plot and also by checking the cook distance and rstandard values..The cookD value was checked using influential plot and the values were not greater than or equal to 1.So there were no influential data points.It is shown in the figure42 and figure43.The next observation was independence of errors.To verify this assumption Durbin-Watson test was used.Here the Durbin-Watson statistic was around 1.7 which is shown in the figure44 This satisfied the assumption of independence of observations.The final assumption was the linearity.Here the independent variables should have a linear relationship between the log-odds of the outcome.So verify this assumption the interaction between the independent variables and the log of itself was included in the glm function.It was observed that all the independent variables were non-statistically significant after running the model.This satisfied the linearity assumption.This is shown in the figure45

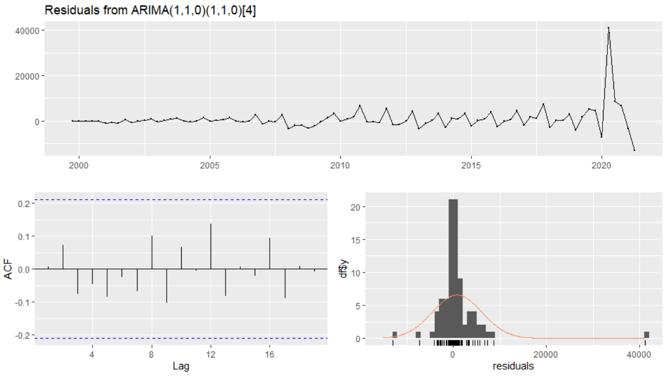


Fig. 26. Residual plots SARIMA model

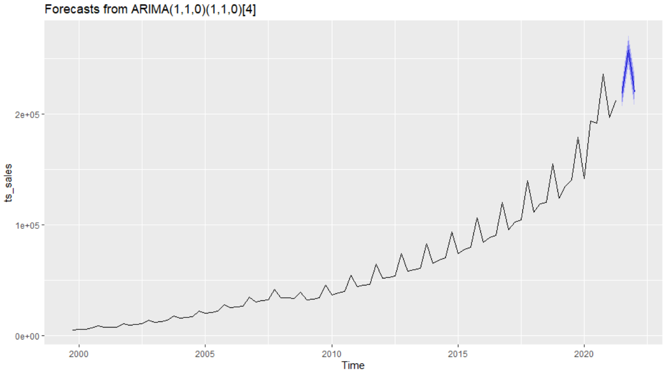
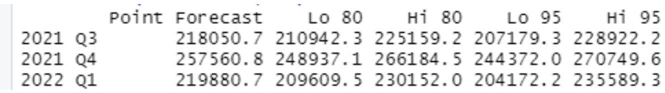


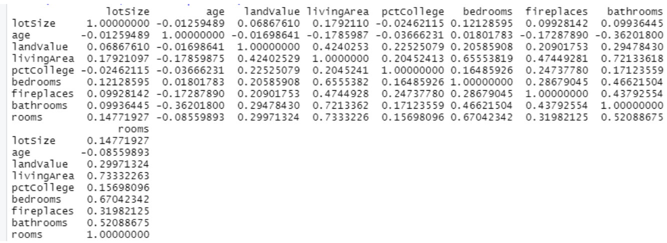
Fig. 27. Plot SARIMA model

The confusion matrix and the other statistics is shown in the figure46. From the confusion matrix it can be seen that the sensitivity rate was 0.7362 and the specificity rate was 0.8573.The kappa value was 0.5981.The accuracy of the model was 0.8022 which is around 80 percent.



```
        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
```

Fig. 28. Forecasts for 3 periods using SARIMA model



```
               lotSize         age   landValue livingArea  pctCollege   bedrooms  fireplaces   bathrooms
lotSize     1.00000000 -0.01259489  0.06867610  0.1792110 -0.02462115 0.12128595  0.09928142  0.09936445
age        -0.01259489  1.00000000 -0.01698641 -0.1785987 -0.03666231 0.01801783 -0.17287890 -0.36201800
landValue   0.06867610 -0.01698641  1.00000000  0.4240253  0.22525079 0.20585908  0.20901753  0.29478430
livingArea  0.17921097 -0.17859875  0.42402529  1.0000000  0.20452413 0.65553819  0.47449281  0.72133618
pctCollege -0.02462115 -0.03666231  0.22525079  0.2045241  1.00000000 0.16485926  0.24737780  0.17123559
bedrooms    0.12128595  0.01801783  0.20585908  0.6555382  0.16485926 1.00000000  0.28679045  0.46621504
fireplaces  0.09928142 -0.17287890  0.20901753  0.4744928  0.24737780 0.28679045  1.00000000  0.43792554
bathrooms   0.09936445 -0.36201800  0.29478430  0.7213362  0.17123559 0.46621504  0.43792554  1.00000000
rooms       0.14771927 -0.08559893  0.29971324  0.7333226  0.15698096 0.67042342  0.31982125  0.52088675
                 rooms
lotSize     0.14771927
age        -0.08559893
landValue   0.29971324
livingArea  0.73332263
pctCollege  0.15698096
bedrooms    0.67042342
fireplaces  0.31982125
bathrooms   0.52088675
rooms       1.00000000
```

Fig. 29. correlation between all the continuous variables

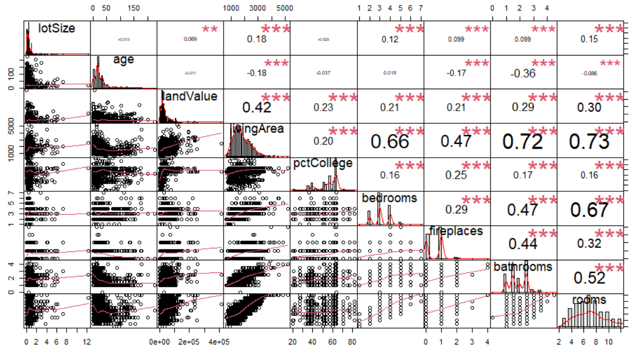


Fig. 30. correlation chart 1

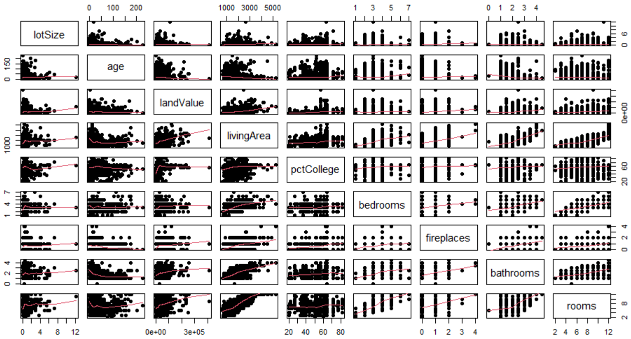
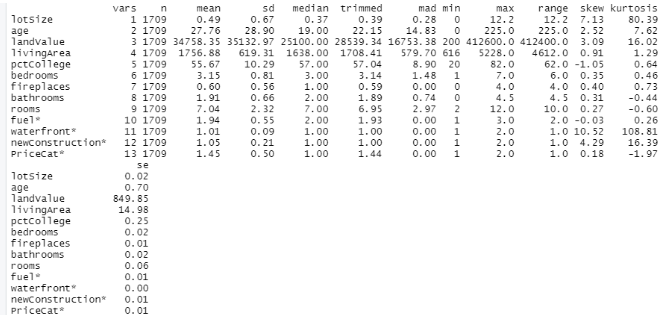
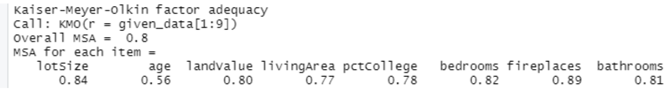


Fig. 31. correlation chart 2



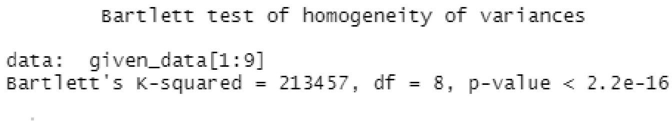
```
                vars    n     mean      sd   median trimmed      mad min      max    range  skew kurtosis
lotSize            1 1709     0.49    0.67     0.37    0.39     0.28   0     12.2     12.2  7.13    80.39
age                2 1709    27.76   28.90    19.00   22.15    14.83   0    225.0    225.0  2.52     7.62
landValue          3 1709 34758.35 35132.97 25100.00 28539.34 16753.38 200 412600.0 412400.0  3.09    16.02
livingArea         4 1709  1756.88  619.31  1638.00 1708.41   579.70 616   5228.0   4612.0  0.91     1.29
pctCollege         5 1709    55.67   10.29    57.00   57.04     8.90  20     82.0     62.0 -1.05     0.64
bedrooms           6 1709     3.15    0.81     3.00    3.14     1.48   1      7.0      6.0  0.35     0.46
fireplaces         7 1709     0.60    0.56     1.00    0.59     0.00   0      4.0      4.0  0.40     0.73
bathrooms          8 1709     1.91    0.66     2.00    1.89     0.74   0      4.5      4.5  0.31    -0.44
rooms              9 1709     7.04    2.32     7.00    6.95     2.97   2     12.0     10.0  0.27    -0.60
fuel*             10 1709     1.94    0.55     2.00    1.93     0.00   1      3.0      2.0 -0.03     0.26
waterfront*       11 1709     1.01    0.09     1.00    1.00     0.00   1      2.0      1.0 10.52   108.81
newConstruction*  12 1709     1.05    0.21     1.00    1.00     0.00   1      2.0      1.0  4.29    16.39
PriceCat*         13 1709     1.45    0.50     1.00    1.44     0.00   1      2.0      1.0  0.18    -1.97
                    se
lotSize           0.02
age               0.70
landValue       849.85
livingArea       14.98
pctCollege        0.25
bedrooms          0.02
fireplaces        0.01
bathrooms         0.02
rooms             0.06
fuel*             0.01
waterfront*       0.00
newConstruction*  0.01
PriceCat*         0.01
```

Fig. 32. Descriptive Statistics



```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = given_data[1:9])
Overall MSA =  0.8
MSA for each item = 
   lotSize        age  landValue livingArea pctCollege   bedrooms fireplaces  bathrooms 
      0.84       0.56       0.80       0.77       0.78       0.82       0.89       0.81 
```

Fig. 33. Summary of KMO test



```
	Bartlett test of homogeneity of variances

data:  given_data[1:9]
Bartlett's K-squared = 213457, df = 8, p-value < 2.2e-16
```

Fig. 34. Summary of Bartlett test

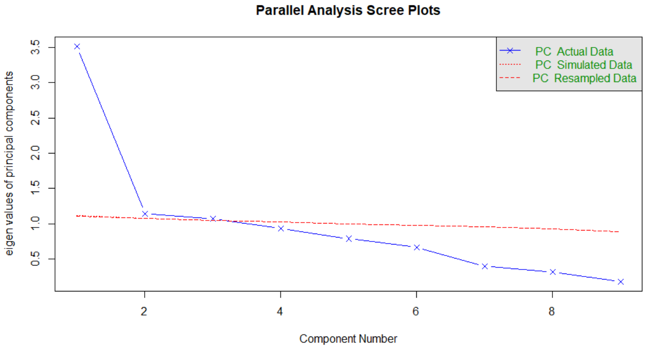
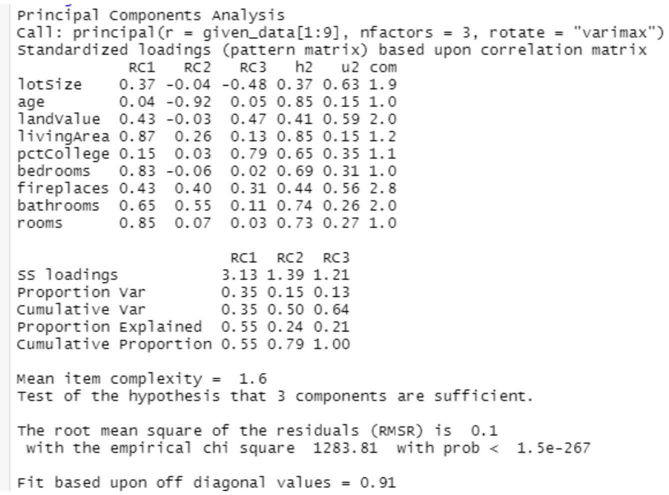


Fig. 35. Scree Plot



```
Principal Components Analysis
Call: principal(r = given_data[1:9], nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
            RC1   RC2   RC3   h2   u2 com
lotSize    0.37 -0.04 -0.48 0.37 0.63 1.9
age        0.04 -0.92  0.05 0.85 0.15 1.0
landValue  0.43 -0.03  0.47 0.41 0.59 2.0
livingArea 0.87  0.26  0.13 0.85 0.15 1.2
pctCollege 0.15  0.03  0.79 0.65 0.35 1.1
bedrooms   0.83 -0.06  0.02 0.69 0.31 1.0
fireplaces 0.43  0.40  0.31 0.44 0.56 2.8
bathrooms  0.65  0.55  0.11 0.74 0.26 2.0
rooms      0.85  0.07  0.03 0.73 0.27 1.0

                       RC1  RC2  RC3
SS loadings           3.13 1.39 1.21
Proportion Var        0.35 0.15 0.13
Cumulative Var        0.35 0.50 0.64
Proportion Explained  0.55 0.24 0.21
Cumulative Proportion 0.55 0.79 1.00

Mean item complexity =  1.6
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.1 
 with the empirical chi square  1283.81  with prob <  1.5e-267 

Fit based upon off diagonal values = 0.91
```

Fig. 36. Summary of PCA

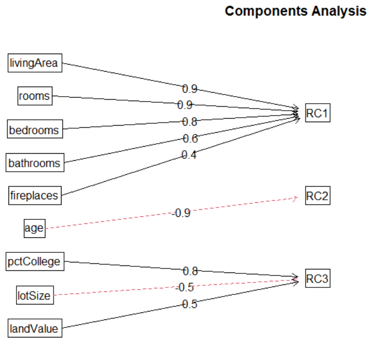
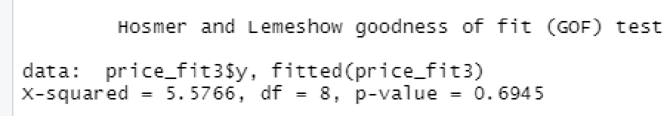
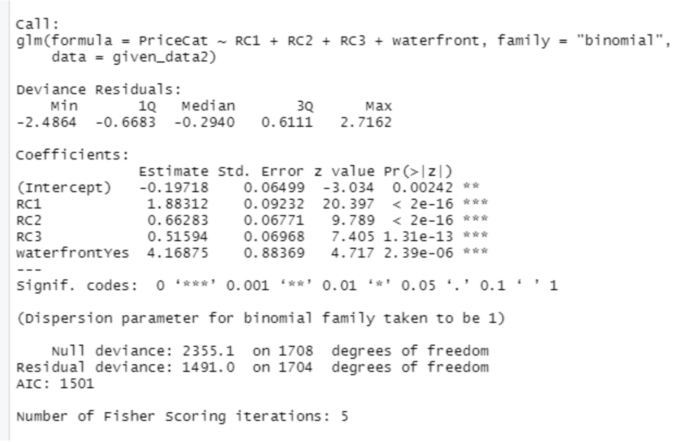


Fig. 37. Components Analysis



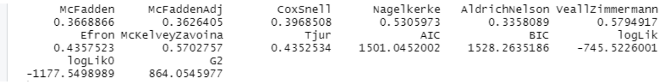
```
	Hosmer and Lemeshow goodness of fit (GOF) test

data:  price_fit3$y, fitted(price_fit3)
X-squared = 5.5766, df = 8, p-value = 0.6945
```

Fig. 38. Hosmer Lemeshow Test for final model



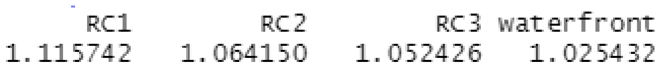
```
Call:
glm(formula = PriceCat ~ RC1 + RC2 + RC3 + waterfront, family = "binomial",
    data = given_data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4864  -0.6683  -0.2940   0.6111   2.7162

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.19718    0.06499  -3.034  0.00242 **
RC1            1.88312    0.09232  20.397  < 2e-16 ***
RC2            0.66283    0.06771   9.789  < 2e-16 ***
RC3            0.51594    0.06968   7.405 1.31e-13 ***
waterfrontYes  4.16875    0.88369   4.717 2.39e-06 ***
---
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: 1491.0  on 1704  degrees of freedom
AIC: 1501

Number of Fisher Scoring iterations: 5
```

Fig. 39. Summary of the final model



```
      McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson VeallZimmermann
     0.3668866       0.3626405       0.3968508       0.5305973       0.3358089       0.5794917
         Efron McKelveyZavoina            Tjur             AIC             BIC          logLik
     0.4357523       0.5702757       0.4352534    1501.0452002    1528.2635186    -745.5226001
       logLik0              G2
 -1177.5498989     864.0545977
```

Fig. 40. Pseudo Rsquare values of various metrics



```
     RC1        RC2        RC3 waterfront
1.115742   1.064150   1.052426   1.025432
```

Fig. 41. Vif values for the independent variables

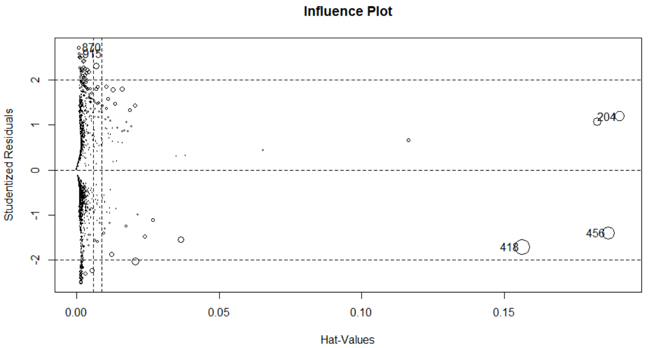
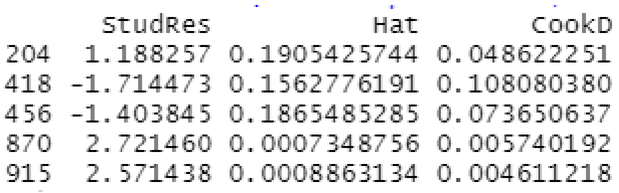
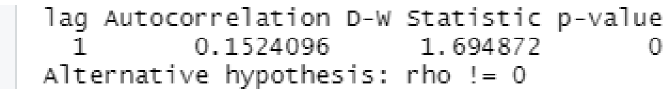


Fig. 42. Influence Plot for the model



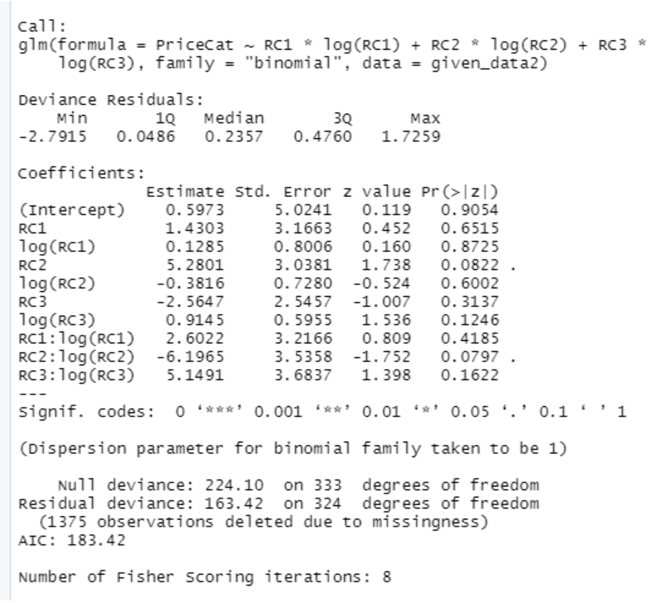
```
      StudRes          Hat        CookD
204  1.188257 0.1905425744 0.048622251
418 -1.714473 0.1562776191 0.108080380
456 -1.403845 0.1865485285 0.073650637
870  2.721460 0.0007348756 0.005740192
915  2.571438 0.0008863134 0.004611218
```

Fig. 43. Cook D values for the indepenent variables



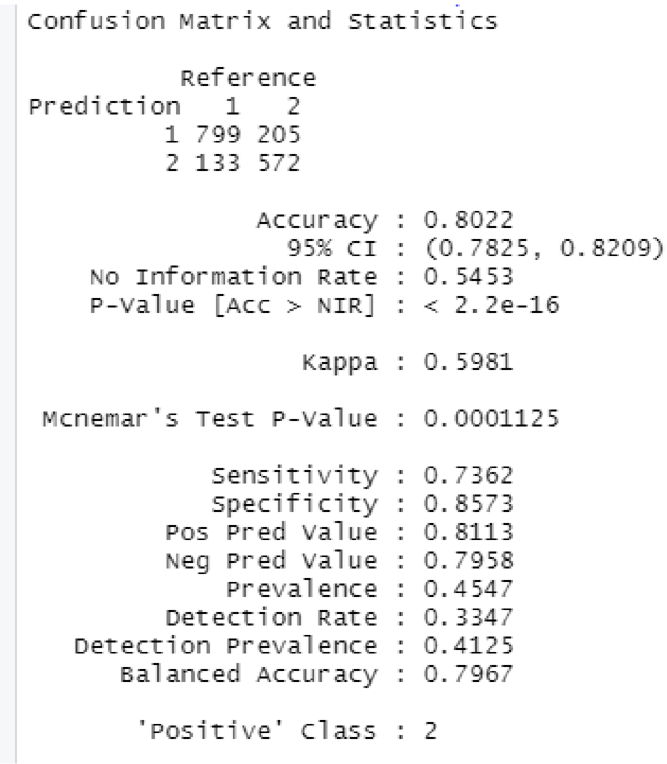
```
 lag Autocorrelation D-W Statistic p-value
   1       0.1524096      1.694872       0
 Alternative hypothesis: rho != 0
```

Fig. 44. Durbin Watson Test



```
Call:
glm(formula = PriceCat ~ RC1 * log(RC1) + RC2 * log(RC2) + RC3 *
    log(RC3), family = "binomial", data = given_data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7915   0.0486   0.2357   0.4760   1.7259

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.5973     5.0241   0.119   0.9054
RC1             1.4303     3.1663   0.452   0.6515
log(RC1)        0.1285     0.8006   0.160   0.8725
RC2             5.2801     3.0381   1.738   0.0822 .
log(RC2)       -0.3816     0.7280  -0.524   0.6002
RC3            -2.5647     2.5457  -1.007   0.3137
log(RC3)        0.9145     0.5955   1.536   0.1246
RC1:log(RC1)    2.6022     3.2166   0.809   0.4185
RC2:log(RC2)   -6.1965     3.5358  -1.752   0.0797 .
RC3:log(RC3)    5.1491     3.6837   1.398   0.1622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 224.10  on 333  degrees of freedom
Residual deviance: 163.42  on 324  degrees of freedom
  (1375 observations deleted due to missingness)
AIC: 183.42

Number of Fisher Scoring iterations: 8
```

Fig. 45. Linearity Assumption



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 799 205
         2 133 572

               Accuracy : 0.8022
                 95% CI : (0.7825, 0.8209)
    No Information Rate : 0.5453
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5981

 Mcnemar's Test P-Value : 0.0001125

            Sensitivity : 0.7362
            Specificity : 0.8573
         Pos Pred Value : 0.8113
         Neg Pred Value : 0.7958
             Prevalence : 0.4547
         Detection Rate : 0.3347
   Detection Prevalence : 0.4125
      Balanced Accuracy : 0.7967

       'Positive' Class : 2
```

Fig. 46. Confusion matrix and other statistics

III. Conclusion

So for the given time series simple time series models : mean model,naive model and seasonal naive model,exponential smoothing models : simple exponential smoothing model,Holt linear exponential smoothing model and Holt Winters exponential model as well as ARIMA and SARIMA models were implemented for forecasting sales.The accuracy was estimated based on the lower RMSE value.Firstly among the simple time series model the lower rmse value was found in seasonal naive model which was 15144.Among exponential smoothing model lower RMSE value was found in Holt Winter Exponential smoothing model which was 4969.59 and its AIC value was 1768.792.Among ARIMA and SARIMA model the lowest RMSE value was found in SARIMA model which was 5318.922 and it's AIC value was 1652.59..So on the basis of lower RMSE value Holt Winters multiplicative model would be the optimal model for the quarterly time series of United States e-commerce retail sales as the series encompasses both seasonal and trend components. For the given data set of US house categories logistic regression model was built to predict the house category which was 'expensive' or 'budget' using the house characteristics.The model was built using two approaches.One using regular approach by including all the independent variables and removing the non-significant variables and another by using principal component analysis method.It was found that the model which was built using the principal component analysis showed a good fit which was examined using Hosmer lemeshow test.It satisfied all the assumptions and the accuracy of the model was 0.8022.The sensitivity rate was 0.7362 and specificity rate was 0.8573.The 'positive' class was 2.

References

[1] "Time series," https://www.investopedia.com/terms/t/timeseries.asp, accessed: 2021-12-29.

[2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.