Statistics for Data Analytics - CA II

*Report for Statistic for Data Analytics course

Juan Pablo Andres Palma Bustos
Master in Science Data Analytics (MSCDAD_B)
National College of Ireland
Dublin, Ireland
x20225776@student.ncirl.ie

***Abstract*—The paper explores the Time Series Analysis and Logistic Regression method applied to two different data frames. The result shows that the suitable Time Series Analysis is a SARIMA model and regular Logistic Regression model in contrast with the model using PCA techniques.**

***Index Terms*—Time Series, SARIMA, PCA, Logistic Regression, E-commerce sales and House pricing Category.**

INTRODUCTION

In this paper, we performed the assignment of the appliances of Time Series Analysis for a data set E-commerce sales in the United States (US) in Billion of dollars using the best performed time series model for the data characteristic.

On the other hand, we applied a Logistic regression model in a data set that wants to classify house prices in Budget or Expensive.

Each section starts with a general description of the data set, followed by the modelling process. Finally, a summary focused on discussing the model's performance, the interpretability of the data, and the respective forecast when possible.

I. PART A - TIME SERIES ANALYSIS

A. Assessment of components

Studies of time-series data aim to answer two fundamental questions. Firstly, the description part tries to explain what happened. Secondly, what will happen next that is called the forecasting part.

As follows in Figure 1, which displays the raw data, it is possible to say two features. Firstly, the data shows a trend pattern increasing over time. Secondly, it is possible to define that the information is not stationary. Determining if the data have a "stationary" component is crucial to declare with model suit better. If the data is not stationary, we must make the difference and model the component.

Also, the time series shows a seasonal pattern, which occurs when a time series is affected by seasonal factors (see Figures 2 and 3).

Fig. 1. Time serie description Raw data

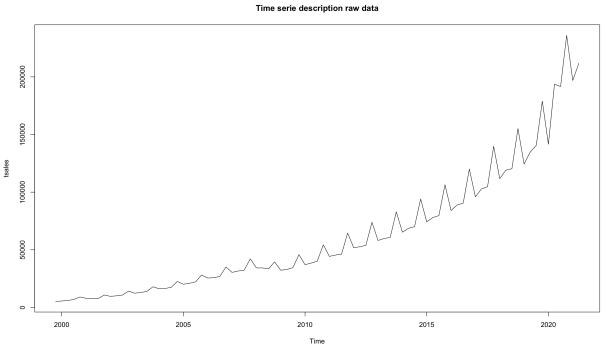


Fig. 2. Seasonal plot of the Time serie

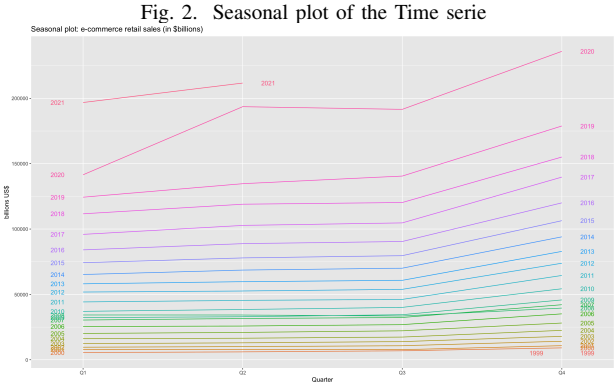
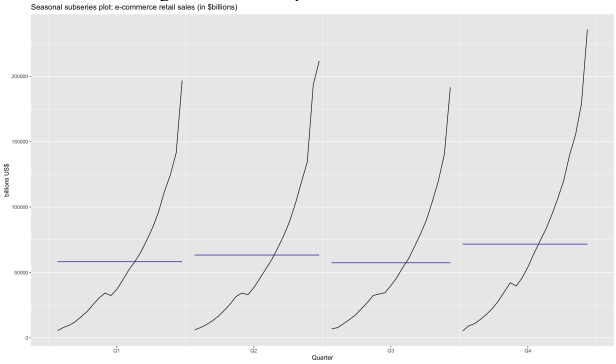


Fig. 3. Subserie plot of the Time serie



B. Modelling

During the modelling process, first, we tried to probe all the models, but the appropriate way was to focus on the internal characteristics of the data.

The fact that data have a trend and seasonal component suggest focusing on the prediction of Holt-Winter's seasonal method the order (M,A,M) or seasonal ARIMA.

Using the package "ets" in R, we compared the model's performance as shown below (See Figure XX).

Fig. 4. Comparison Triple

Next step, it is necessary to compare now the performance of the model proposed by R in the last Figure with the Seasonal ARIMA model.

Firstly, we have to prove statistically that the data is not stationary. We used the Dickey-Fuller Test, which expected the result to be non-significative (see Figure 5).

Fig. 5. Dickey-Fuller Test for check seasonality

```
        Augmented Dickey-Fuller Test

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

We will continue with the modelling of the SARIMA model. Firstly, we display a comparative between the E-commerce sales data and its logarithm (See Figure 4).

C. Summary

The best SARIMA model shows to be the order (1,1,0)(1,1,0)[4], which means an auto-regressive element of order 1, moving average 1, and difference of 1, and the respective seasonal part of (1,1,0), as well.

Fig. 6. Seasonal ARIMA model

```
Series: tsales
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 784.1103 5318.922 2390.89 0.291582 3.651811 0.2408179 0.006567588
```

Fig. 7. Q-Q Plot

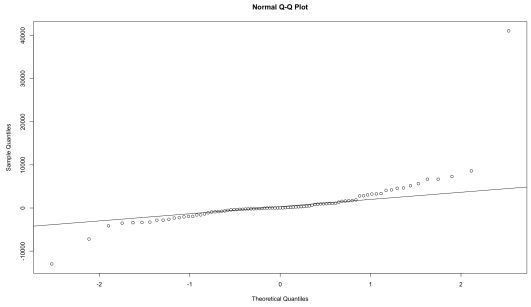
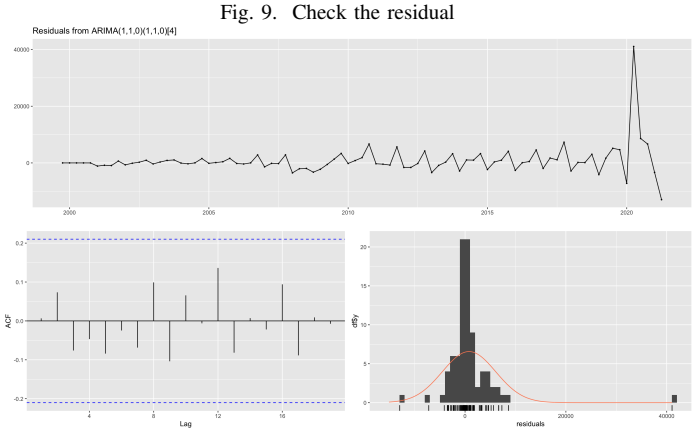


Fig. 8. Box-Ljung test

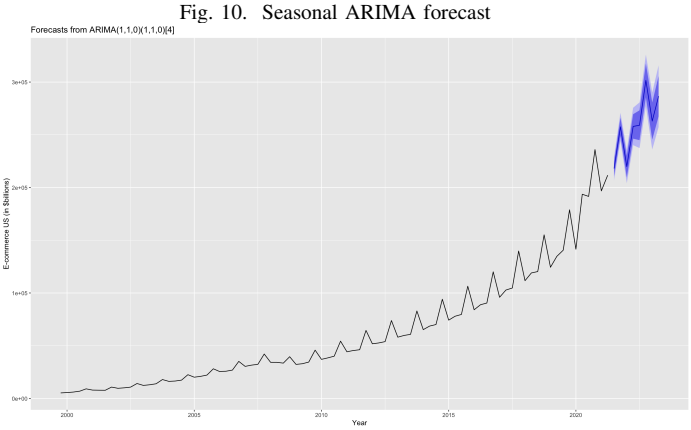
```
        Box-Ljung test

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

Fig. 9. Check the residual



Also, this model presents the lowest RMSE and AIC as criteria of evaluation. Figure 10, show the final result of the forecast using the SARIMA model. It performs well in the Box-Ljung test and present good fit in the residual plot (see Figure 9), and a very plausible forecast.

Fig. 10. Seasonal ARIMA forecast



II. PART B - Logistic Regression

A. Descriptive Statistics

In an overview, we Notice that we have nine numeric variables (see Table II) and four categorical (see Table I).

The dependent variable (target variable) is the house's price category like Budget or Expensive, in a proportion of 55% and 45%, respectively, in the database.

Moreover, the database has a high proportion of houses that are not a new construction (95% olders), being the gas the most predominant mechanism for heating (69%), electric (18%), and oil (12%). And a 99% have not WaterFront.

TABLE I
STATISTICAL SUMMARY

		Price Category			
		Budget N	Expensive N	Budget %	Expensive %
Fuel	electric	243	69	0.14	0.04
	gas	553	634	0.32	0.37
	oil	136	74	0.08	0.04
WaterFront	No	930	764	0.54	0.45
	Yes	2	13	0.00	0.01
New Construction	No	920	709	0.54	0.41
	Yes	12	68	0.01	0.04
Price Category	Budget	932	-	0.55	-
	Expensive	-	777	-	0.45

On the other hand, the database has the lotsize with a mean of 0.49 and maximum of 12, age of the house in average is 28 years old, but it also have old houses with a maximum of 225 years old. Also, we have the landValue, Living area, percentage of local people with college education with a mean of 55%. So it suggests that the houses generally have a good level of education in the neighbourhood. Finally, the data set has the number of rooms, bedrooms, fireplace and bathrooms. We just noticed that number of bathrooms has no integer values. So it will be assumed as a toilet with no shower or something like this.

Even when the database does not present missing values, it has some nonsense or unlikely data, like a house with no toilets or lotSize of 0. So we recoded these issues using the following lower value in the data frame.

TABLE II
STATISTICAL SUMMARY NUMERICAL VARIABLES

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
lotSize	1,709	0.49	0.67	0.01	0.17	0.54	12.20
age	1,709	27.76	28.90	0	13	34	225
landValue	1,709	34,758	35,133	200	15,100	40,500	412,600
livingArea	1,709	1,757	619	616	1,302	2,144	5,228
pctCollege	1,709	55.67	10.29	20	52	64	82
bedrooms	1,709	3.15	0.81	1	3	4	7
fireplaces	1,709	0.60	0.56	0	0	1	4
bathrooms	1,709	1.91	0.66	1	1.5	2.5	4
rooms	1,709	7.04	2.32	2	5	9	12

Furthermore, in Figure 11 is possible to see the Partial correlations between numerical variables. The higher correlation between bathrooms and living area (0.72), or living area and room, and bedroom, 0.73 and 0.66, respectively. We have to look after these variables during the modelling process because they can produce problems of multicollinearity or because these sets of variables with higher correlations could be a good candidate for testing dimension reduction methods.

Fig. 11. Partial Correlations categorical variables

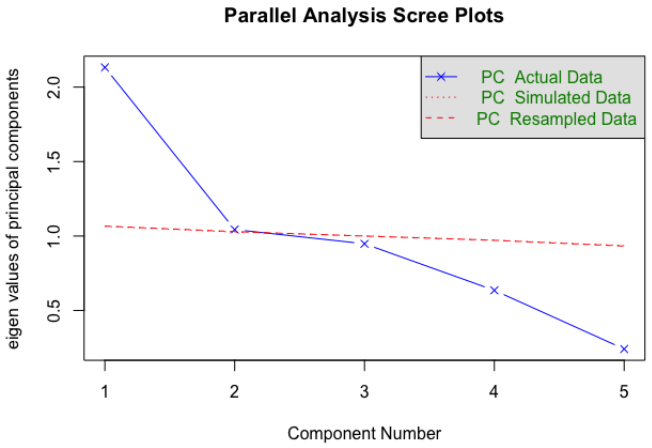
B. Principal Component Analysis (PCA)

In the House topic, it is plausible to think that the house price is determined for internal characteristics such as the

number of rooms, toilets or others, and other external factors like the neighbourhood or the material construction or the size itself. In any case, the idea of putting together some variables that can jointly determine the price of the house is a hypothesis reasonable and worthy to try some techniques like Principal Component Analysis.

As shown in Figure 12, the Scree plot recommend two factors, and in an iterative process, we combine all the variables and select which ones had higher loads in the components.

Fig. 12. PCA Scree Plot



After combining many alternatives, we test the PCA with two factors and the subset shown in Figure XX, where livingArea, rooms, bedrooms and bathrooms make one component. And a second component with pctCollege, landValue and Fireplaces.

Fig. 13. PCA loads and Components

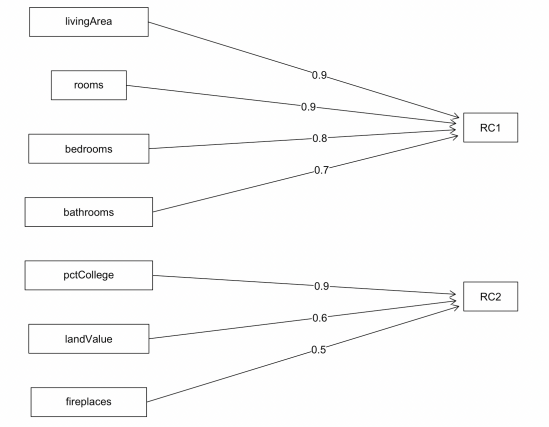


TABLE III
NULL MODEL VS. FULL MODEL

	Model 0	Model 1
(Intercept)	-0.18^{***}	-6.81^{***}
	(0.05)	(0.54)
lotSize		0.53^{***}
		(0.15)
age		-0.01
		(0.00)
landValue		0.00^{***}
		(0.00)
livingArea		0.00^{***}
		(0.00)
pctCollege		-0.02
		(0.01)
bedrooms		-0.11
		(0.13)
fireplaces		0.04
		(0.15)
bathrooms		1.00^{***}
		(0.16)
rooms		0.02
		(0.05)
fuelgas		0.22
		(0.21)
fueloil		0.03
		(0.29)
waterfrontYes		3.40^{***}
		(0.95)
newConstructionYes		-0.26
		(0.44)
AIC	2357.10	1311.40
BIC	2362.54	1387.62
Log Likelihood	-1177.55	-641.70
Deviance	2355.10	1283.40
Num. obs.	1709	1709

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

C. Modelling

In the modelling section, we focused on the logistic regression without PCA. It is because selecting the better variables does not fit with the logic of dimension reduction. However, in the final Table, we will compare the model of both alternatives and their respective performance.

Table III, shows the comparison between the Null model (no variables) against the Full model. The result is that the set of predictive variables significantly affects predicting into classification task. However, not every variable contribute the same.

Displayed the Full model, we can consider which variables are not significantly contributing to the model. Still, it also will be supported by the omnibus test (see Table IV). The result of the process is a subset of variables where lotSize, age, landValue, livingArea, pctCollege, bathrooms and waterFront are the most contributive variables in the model performance.

Models 2, 3 and 4 in Table V show that the final model discarded the non-significant variables that suggest the omnibus test. The Table also compares the model using the PCA components.

The parsimony criteria in this context are open to discussion in an overview. Firstly, the assignment criteria such as AIC and BIC are similar. However, the logistic regression in

TABLE IV
OMNIBUS TEST

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			1708.00	2355.10	
lotSize	1.00	61.82	1707.00	2293.28	0.00
age	1.00	43.87	1706.00	2249.40	0.00
landValue	1.00	424.29	1705.00	1825.11	0.00
livingArea	1.00	476.93	1704.00	1348.18	0.00
pctCollege	1.00	3.95	1703.00	1344.23	0.05
bedrooms	1.00	0.35	1702.00	1343.88	0.56
fireplaces	1.00	2.31	1701.00	1341.57	0.13
bathrooms	1.00	41.39	1700.00	1300.18	0.00
rooms	1.00	0.10	1699.00	1300.09	0.76
fuel	2.00	0.93	1697.00	1299.16	0.63
waterfront	1.00	15.43	1696.00	1283.73	0.00
newConstruction	1.00	0.33	1695.00	1283.40	0.57

TABLE V
MODEL BUILDING AND PARSIMONY

	Model 2	Model 3	Model 4	Model 5 (PCA)
(Intercept)	-6.91^{***}	-7.53^{***}	-7.12^{***}	-0.29^{**}
	(0.52)	(0.38)	(0.41)	(0.11)
lotSize	0.49^{***}	0.52^{***}	0.51^{***}	0.47^{***}
	(0.14)	(0.14)	(0.14)	(0.12)
age	-0.01^{*}	-0.01^{*}	-0.02^{**}	-0.01^{*}
	(0.00)	(0.00)	(0.01)	(0.00)
landValue	0.00^{***}	0.00^{***}	0.00^{***}	
	(0.00)	(0.00)	(0.00)	
livingArea	0.00^{***}	0.00^{***}	0.00^{***}	
	(0.00)	(0.00)	(0.00)	
pctCollege	-0.01			
	(0.01)			
bathrooms	1.01^{***}	1.00^{***}	0.92^{***}	
	(0.16)	(0.16)	(0.16)	
waterfrontYes	3.41^{***}	3.58^{***}	3.66^{***}	4.21^{***}
	(0.94)	(0.93)	(0.94)	(0.88)
age^2			0.00^{*}	
			(0.00)	
RC1				1.71^{***}
				(0.09)
RC2				1.02^{***}
				(0.08)
AIC	1301.86	1302.77	1299.36	1494.39
BIC	1345.41	1340.87	1342.91	1527.05
Log Likelihood	-642.93	-644.38	-641.68	-741.20
Deviance	1285.86	1288.77	1283.36	1482.39
Num. obs.	1709	1709	1709	1709

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

model 3 will perform better.

Comparing the two better performance models (model 3 and model 5) is small (see Tables VI and VII). However, the accuracy of model 3 (Regression model) is 0.84 (0.835), against the accuracy of model 5 0.81 (0.805). Model 3 performs better in general.

We also tried a model with only significant variables and applied PCA to the subset. The improved PCA was closer to model 3 performance, but doing that exercise, the reason for doing PCA were weak, even when the result was getting close to model 3.

TABLE VI
CONFUSION MATRIX LOGISTIC REGRESSION

	FALSE	TRUE
Budget	819	113
Expensive	169	608
Accuracy	**0.835**	

TABLE VII
CONFUSION MATRIX LOGISTIC REGRESSION WITH PCA

	FALSE	TRUE
Budget	801	131
Expensive	202	575
Accuracy	**0.805**	

D. Summary

Finally, we must evaluate if model 3 respects all the assumptions necessaries for Logistic regression.

Firstly, the vif test for checking multicollinearity. Every variable considered in the model has a tiny number, which is strong evidence to discard multicollinearity problems (see Figure 14).

Fig. 14. Check for Multicollinearity

```
vif(housefit14)
  lotSize       age  landValue livingArea  bathrooms waterfront
1.023860  1.268526   1.123492   1.365316   1.560952   1.018913
```

In Figure 15, we checked the linearity of the predicted values, and even when they are not perfectly fitted, it is impossible to see significant issues on this assumption.

Fig. 15. Check for Linearity of the predicted values

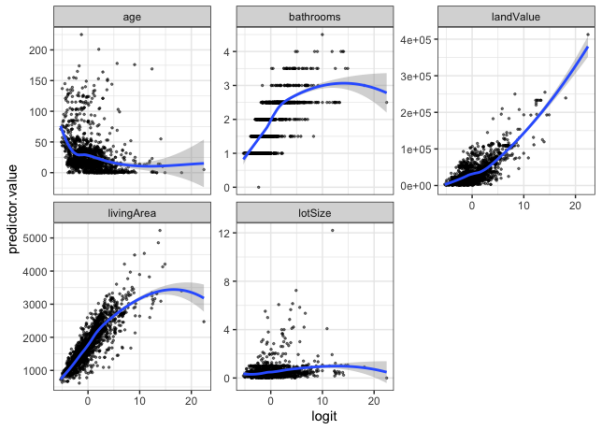
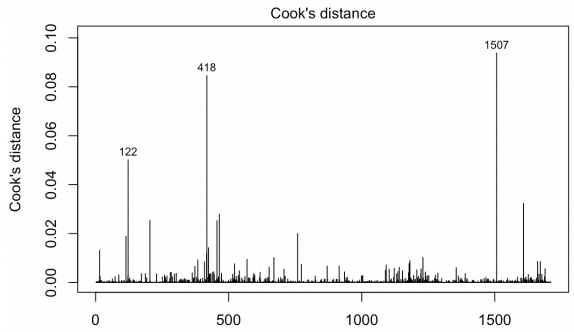


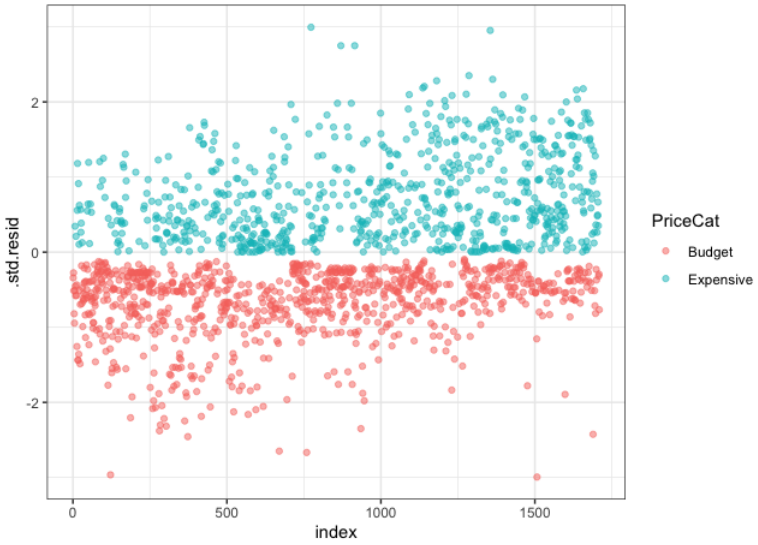
Fig. 16. Cook distance



On the other hand, we check the cook distance to identify cases that affect the model's performance. The test identified three instances, but the difference was insignificant when reviewing the performance with and without them (see 16).

Another significant assumption of the model is to check the distribution of the residual, as it is shown in Figure 17 there are not a pattern that can explain the distribution of the residual, meeting the assumption for a good fit of the model.

Fig. 17. Distribution of the residuals by category Dependent Variable



Finally, in terms of the interpretability of the results, it is possible to say that the odds ratio of having a waterfront are 35.71 times the odds of not having one. In the same way, bathrooms, livingArea, landValue and Age have a positive effect. However, only lotSize has a negative impact where the increase of 1 unit means 0.99 times the odds.

TABLE VIII
COEFFICIENT AND INTERPRETATION

	coef	exp(coef)
(Intercept)	-7.5401	0.0005
lotSize	-0.0062	0.9938
age	4.2005e-05	1.0000
landValue	0.0023	1.0023
livingArea	0.5185	1.6795
bathrooms	1.0009	2.7207
waterfrontYes	3.5754	35.7071

REFERENCES

[1] R. Hyndman, G. Athanasopoulos. 2018. Forecasting: Principles and Practice. https://otexts.com/fpp2/

[2] A. Kassambara. 2017. Machine Learning Essentials: Practical Guide in R.

[3] K. Leung. 2021. Assumptions of Logistic Regression, Clearly Explained. https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290