National College of Ireland

Project Submission Sheet – 2021/2022

Student Name: …Polina Prinii……………………………………………………………………………………

Student ID: …x21137757……………………………………………………………………………………..

Programme: …PGDDA……………………………………………… **Year:** ………2021………

Module: …Statistics for Data Analytics……………………………………………………………

Lecturer: …Hicham Rifai…………………………………………………………………………….

Submission Due Date: …10/01/2022…………………………………………………………………………….

Project Title: … Part A: Time Series Analysis....&....Part B: Logistics Regression Analysis...................

Word Count: …6102…………………………………………………………………………………………….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: …Polina Prinii……………………………………………………………………………

Date: …30/12/2021…………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Part A: Time Series Analysis

Statistics for Data Analytics – Terminal Assignment-Based Assessment

Polina, Prinii
School of Computing
National College Ireland, Mayor Street Lower, IFSC, Dublin 1, D01 Y300, Ireland.
x21137757@student.ncirl.ie

***Abstract* — This document serves the purpose of a findings report in a business format. The aim is to outline the data, methods applied and findings when performing and reporting on a Time Series analysis with the goal of prediction.**

I. Introduction

The United States is home to over a million retail establishments raking in at 1,045,422 [1] as of the year 2020, generating a whopping 5.15 trillion US dollars in 2021 [2] for the American economy. Being as the retail sector is a major financial factor to the American economy, the purpose of this analysis is to analyze the eComm_US.csv file, a compilation of quarterly retail sales (in billions) commencing from Q4 1999 with the aim of forecasting Q3, Q4 of 2021 and Q1 of 2022.

Objective:

Using Time Series Analysis, the objective of this report is to forecast the three periods ahead to produce statistically valuable information.
A time series is a sequence of observations taken at successively equal spaced-out points of time [3]. A time series analysis is a specific method of analyzing time series with the goal of observing how the variables change over time and derive valuable insights [4].
The following analysis is undertaken using the R scripting language, facilitated through RStudio.

II. Methodology

The analysis aims to perform the following methods to support the building of the best fit model for the time series analysis with the aim of reporting on the results and findings:

A. Preliminary Assessment:

Preliminary assessment is undertaken to obtain an understanding of the nature and components of the raw time series which is the eComm_US.csv file. Preliminary analysis and the assessment of such helps identify the following:

1. Pattern of the time- series (Horizontal aka Stationary, Trending, Seasonal and or Cyclical)
2. Irregular roughness (dampened by Smoothing)
3. Seasonal Decomposition

The above is constructed using the several libraries and functions available in R.

B. Model Identification:

This section looks to estimate and discuss the identification of a suitable forecasting time series model. There are multiple models which can be applied to time series analysis. The analysis applied the following time-series models:

- Simple Time Series Models
- Exponential Smoothing Models
- ARIMA (Auto Regressive Integrated Moving Average) / SARIMA (Seasonal Auto Regressive Integrated Moving Average) Models

The optimum model with the best fit for the time series under analysis has been chosen based on findings and the decision for said model is discussed throughout the “Result & Evaluations” chapter.

The above is constructed using the several libraries and functions available in R.

III. Results & Evaluations

A. *Preliminary Assessment Findings:*

On the question of the preliminary analysis of the time series in question a clear pattern for the data has been identified.
A display of a trend in the form of a gradual increase is clear. Figure 1 outlines the strong and positive incline from the start to the end of the time series.

Further analysis through the support of the Smoothing process [5] allows for the time series to address any significant irregular or error components it may present. The damping of the irregular/error components is achieved using the simplest methods: simple moving averages which is denoted by the following equation:

$$S_t = (Y_{t-q} + ... + Y_t + ... + Y_{t+q}) / (2_q + 1)$$

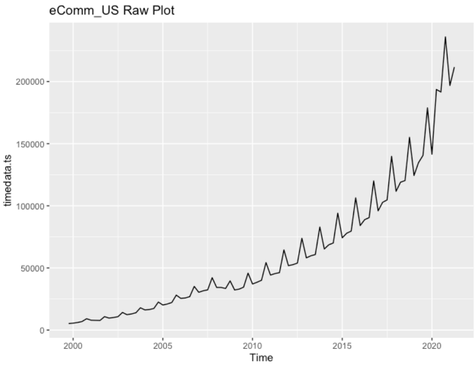


Figure 1: eComm_US Raw Plot

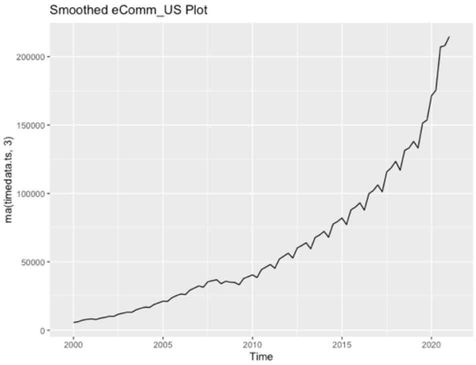


Figure 2: Smoothed eComm_US Plot

Figure 2 confirms the assumption that the data may be of seasonal pattern, as well as trend pattern. Though, very faint the analysis can identify a seasonal pattern commencing from roughly 2001, becoming more predominant roughly in 2009.

To support the pattern of seasonality, a seasonal plot was drafted against the time series object. Figure 3 clearly demonstrates that Q4 is a peak season for retail for all years, though fair prior to 2010 seasonality is evident from the first observation.

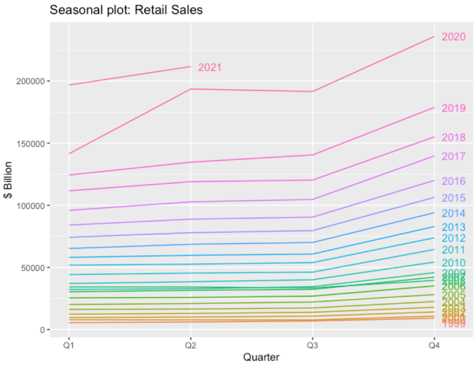


Figure 3: Seasonal plot: Retail Sales

Concluding the preliminary assessment, the time series was decomposed. Seasonal decomposition allowed for the time series to be decomposed into three components, being: -trend, -seasonal and -irregular (or error). Thus, providing an insightful abstract model as illustrated by Figure 4 which allows for the best forecasting method selection to occur. Seasonal decomposition occurred on a multiplicative model and is denoted by:

$$Y_t = Trend * Seasonal_t * Irregular_t$$

In conclusion, Figure 4 confirms and strengthens the understanding around the pattern of the time series, with confidence the analysis can deduct that both a seasonal and trend pattern is present.

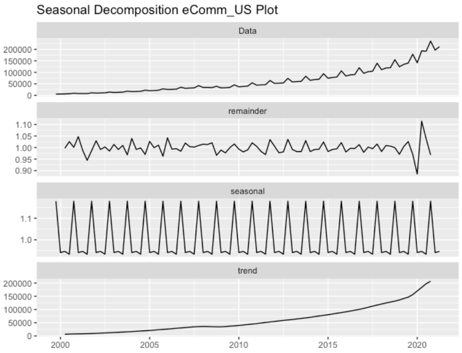


Figure 4: Seasonal Decomposition Plot

B. *Model Identification Findings:*

1) *Simple Time Series Models*

Some forecasting methods can be extremely simple to perform, the simplest being the Mean Forecasting otherwise known as the Average Method. For the purposes of the time series analysis in question the Average Method is not performed on the account of trend and cyclical effect not being considered in the value of the forecast [6].

Thus, the analysis looked at more appropriate simple models such as: - Naïve (Random Walk), - Seasonal Naïve and -Drift method.
The selection for the best model was determined based on the results of the forecast "error" (not be confused with error interpreted as mistake), the error being the difference between the observed value and the forecast [7]. Summarisation of forecast errors aided in the measurement of the forecast accuracy.

Figure 5 outlines the forecasting error measures; the report focuses on interpreting the RMSE results as this indicates the absolute fit of the model to the data. In other words, the RMSE (Root-Mean-Squared-Error) explains in a numerical value how close the value which has been observed is to the model's forecasted value [8].

Model:	ME	RMSE	MAE	MPE	MAPE	MASE
Naïve	***2400.733***	***15390.34***	***9385.384***	***2.855386***	***13.10789***	***0.9453254***
Seasonal Naïve	***9786.711***	***15144***	***9928.205***	***15.44986***	***15.85113***	***1***
Drift	***-2.199491e-12***	***15201.94***	***9399.18***	***-6.090458***	***15.59394***	***0.946715***

Figure 5: Forecasting Error Measures

To conclude, the report with confidence can determined that the Seasonal Naïve Model is the best fit for the time series analysis at the current time, as the RMSE states that a 15,144 deviation is currently present between the observed and forecasted value. This is considerably lower to the Naïve and Drift models. The report now moves to evaluate Exponential Smoothing Model selection.
The report now moves to evaluate Exponential Smoothing model selection.

2) *Exponential Smoothing Model Findings:*

Exponential Smoothing is a powerful forecasting method, like some of the Simple models such as Naïve, the method is used as an alternative to popular Box-Jenkins ARIMA methods.
Like the Naïve Model method, Exponential Smoothing applies a weighted sum of past observation however, the model explicitly applies a decreasing weight for past observations.

There are three models which fall under Exponential Smoothing:

- Simple Exponential Smoothing
- Holt's Linear
- Holt-Winters Seasonal

Having identified trend patterns and seasonality as part of preliminary assessment, the analysis conducted a Holt-Winters Seasonal method. The Holt-Winter method was applied to the time series on a both additive and multiplicative variation. Considering the results from both variations it was to no surprise that the multiplicative had the best fit with an RMSE value of 4969.59. The multiplicative method is preferred to when the seasonal variations change to the proportions of the level of the series [9]. Figure 6 pools both variations against the time series, outlines the forecasts.

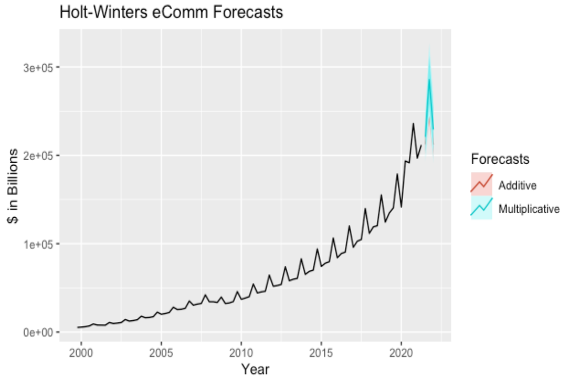
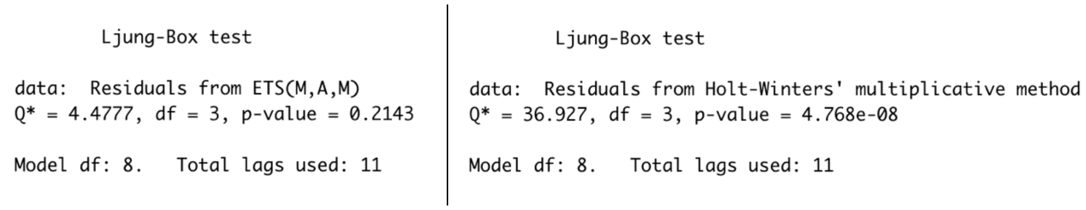


Figure 6: Holt-Winters Forecasts

Comparing the results from the Holt-Winter model to the Seasonal Naïve model the report can move to concluding that the current best fit mode for forecast to the time series at hand is the Holt-Winters. This is supported by the reduction of the RMSE measure from 15,144 (Seasonal Naïve) to 4969 in Holt-Winters.

Software can assist with the selection of the best fit model from exponential smoothing. As the report has found that the Holt-Winters model fits the time series best through manual selection, the report looked at fitting the best fit model automatically. This was possible through the application of the **ets()** function in R. Though the automatic selection resulted with a different RMSE value of 5397.091 this was to be expected.

The RMSE value derived automatically and its difference to the Holt-Winters model performed manual can be explained by the functions in R using different optimization routines as well as different starting values [10]. From Figure 7 it can be conclude that the best fit model is derived from the automatic selection rather the manual, the "Ljung-Box Test" for the EST model [11] confirms this based on the smallest p-value present. Though, the report takes into consideration that the ETS (M, A, M) selection is equivalent to the Holt-Winters multiplicative model.



```
        Ljung-Box test

data:  Residuals from ETS(M,A,M)
Q* = 4.4777, df = 3, p-value = 0.2143

Model df: 8.   Total lags used: 11
```

```
        Ljung-Box test

data:  Residuals from Holt-Winters' multiplicative method
Q* = 36.927, df = 3, p-value = 4.768e-08

Model df: 8.   Total lags used: 11
```

Figure 7: Ljung-Box Test for ETS(MAM) & Holt-Winters Multiplicative Method

The report now moves to evaluate ARIMA/SARIMA model selection.

3) *ARIMA / SARIMA Findings:*

ARIMA or in other words Auto Regressive Integrated Moving Average is a class of models that can explain a time series under analysis based on the past observations of the said time series. Using the lags and the lagged forecast errors an equation can be applied to predict future values [12], the equation denoted as:

$$y_t' = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + {}^{\theta}1\varepsilon_{t-1} + \ldots + {}^{\theta}{}_q\varepsilon_{t-q} + \varepsilon_{t'}$$

ARIMA models best fit stationary time series data, to be more specific time series data which are free of trend and seasonality. SARIMA or Seasonal Auto Regressive Integrated Moving Average models are designed to work with time series data which display seasonality as well as trend.

As mentioned ARIMA is a class of models, thus far the report has evaluated two ARIMA models being:

- Random Forest Walk equivalent to ARIMA(0,1,0)
- Seasonal Naïve equivalent to ARIMA(0,0,0)(0,1,0)

Both ARIMA models showed fairly good results however, the Exponential Smoothing model (M,A,M) equivalent to Holt-Winters has demonstrated best fit model for forecasting future values.

First step to applying a non-stationary time series to an ARIMA model is to plot the raw data and identify the number of differences required to transform the data to stationary.

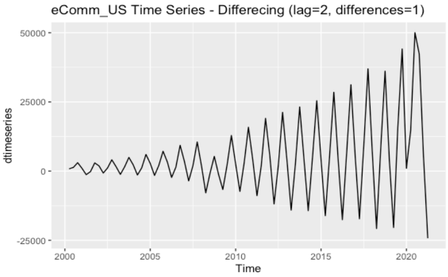


Figure 8: Application of Differencing

Differencing is a method often used to stabilize the mean of a time series thus eliminating trend and seasonality, this is achieved by removing the changes in the level of a time series [13]. Figure 1 in the Preliminary Assessment shows the raw time series plotted, whilst Figure 8 demonstrates the change from non-stationary to stationary using Differencing.

Additionally, Figure 9 outlines the before and after of both the AC (Auto Correlation) function and PAC (Partial Auto Correlation) function. The results conclude the ARIMA model fitted as follows, denoted as (p, d, q):

- p equating to 0. Explained by plot d) as the PACF is gradually decreasing to 0 as the lags increase.
- d equating to 1. Explained by Figure 8 as the analysis derived 1 level of differences.
- q equating to 1.Explained by plot b) as the ACF shows only 1 spike prior to the data returing to a normal level (within the dotted line).

Thus the ARIMA model applied to the time series under analysis shows as (0, 1, 1).

Before proceeding with the model applied, the analysis looks to evaluate the fitness of the models by performing a series of residuals checks. The aim here is for the residuals to effectively be “white noise” and normally distributed.

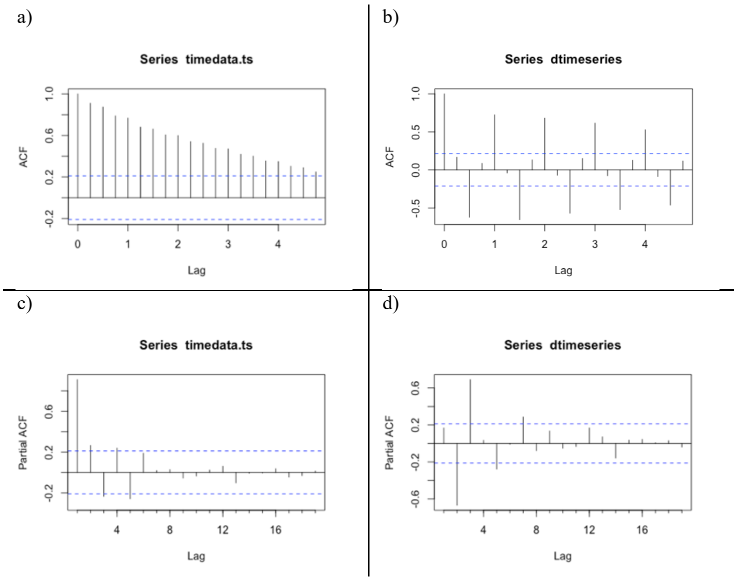


Figure 9: ACF & PACF Plots

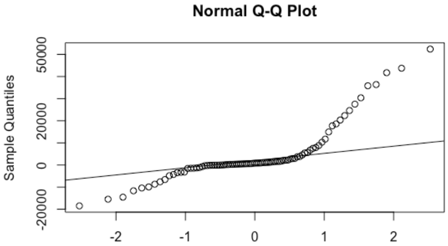
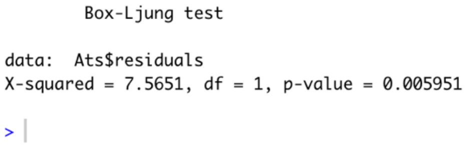


Figure 10: eComm_US Normal Q-Q Plot

Figure 10 outlines a Normal QQ plot with a reasonably normal distribution. However, the decided factor lies within the p-value derived by the “Ljung-Box Test” applied on the ARIMA model. Figure 11 indeed demonstartes that the p-value is of no great significance thus concluding that the applied ARIMA model is of good fit to the time series under analysis.



```
	Box-Ljung test

data:  Ats$residuals
X-squared = 7.5651, df = 1, p-value = 0.005951

>
```

Figure 11: Ljung-Box Test - Testing ARIMA Residuals

For comparison of best fit models for the time series at hand, the analysis looked to software to automatically assign an ARIMA model best fit for the time series. This step was conducted with the question in mind of, if a smaller AIC would be derived through auto.arima against the AIC derived from manula transformation and selection of the appropriate ARIMA model. AIC which stands for Akaike Information Criterion is an additional estimator of prediction widely used when evaluating model fitness.

Auto ARIMA selection results were as to be expected, with the selection of a SARIMA for the time series under analysis as the data display trend and seasonality patters. Please note the analyses aided the auto selection by setting the stepwise to FALSE, which in simpler terms allows for a more robust and thorough search for the best fit ARIMA.

The SARIMA denoted as: ARIMA (1, 1, 0) (1, 1, 0) derived an AIC value of 1646.99 against the manual ARIMA result of 1881.12. A relatively small difference though, for optimal best fit for forecasting the analysis concludes that the auto ARIMA selection is of better fit.

The next section of the report looks to conclude the choice of the most optimum model for the eComm_US time series data with the aim of producing three future forecast periods.

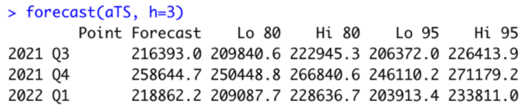
IV. Conclusion

Overall, the results suggest the model selection within the same class are of good fit for the time series under analysis. The report will turn to the evaluation of the RMSE value derived from each optimum model within its class. Though, the AIC value is a good predictor of fitness, it cannot be used for selection between model classes.

Model Name:	Class:	RMSE Value:
Seasonal Naïve	Simple	15144
Holt-Winter (Auto)	Exponential Smoothing	5397.091
ARIMA (1, 1, 0) (1, 1, 0)	SARIMA	4829.919

Figure 12: RMSE Results

Figure 12 outlines the derived RMSE for the three chosen models. The selection for the best fit model is quite evident, thus the analysis conducted the forecast for the three future periods based on the SARIMA models. Figure 13 shows the predicted figures.



```
> forecast(aTS, h=3)
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2021 Q3       216393.0 209840.6 222945.3 206372.0 226413.9
2021 Q4       258644.7 250448.8 266840.6 246110.2 271179.2
2022 Q1       218862.2 209087.7 228636.7 203913.4 233811.0
```

Figure 13: ARIMA Forecast Results

Though the difference between the Holt-Winter and ARIMA models is relatively small, for simplicity reasons the report can take the approach of forecasting using Exponential Smoothing as the model is easier to interpret.

V. Acknowlodgments

I would like to express my appreciation to National College of Ireland who has provided me the possibility to complete this report. A special thanks I give to our lecturer Hicham Rifai who teaches Statistics for Data Analytics, as well as Tony Delaney for producing the course material online. Both have helped me to coordinate my project, especially in understanding the objective and in writing this report.

VI. References

[1] National Retail Federation, "State of Retail", Accessed on: Dec 23rd 2021, [Online], Available at: https://nrf.com/topics/economy/state-retail

[2] Statista Research Department, "United States: total retail sales 2019 – 2025", 2021 Jul 13th, Accessed on Dec 23rd 2021, [Online], Available at: https://www.statista.com/statistics/443495/total-us-retail-sales/

[3] Wikipedia, "Time Series", 2021 Nov 30th, Accessed on Dec 23rd 2021, [Online], Available at: https://en.wikipedia.org/wiki/Time_series

[4] Tableau, "Time Series Analysis: Definition, Types, Techniques, and When It's Used", Accessed on Dec 23rd 2021, [Online], Available at: https://www.tableau.com/learn/articles/time-series-analysis

[5] Adam Hayes, "Simple Moving Averages (SMA)" , 2021 Dec 9th, Accessed on Dec 25th 2021, [Online], Available at: https://www.investopedia.com/terms/s/sma.asp

[6] The Institute of Chartered Accountants of India, "Time Series Analysis & Forecasting", Accessed on Dec 25th 2021, [Online], Available at: https://www.kluniversity.in/arp/uploads/2093.pdf

[7] otexts, "Evaluating forecast accuracy", Accessed on Dec 26th 2021, [Online], Available at: https://otexts.com/fpp2/accuracy.html

[8] Wikipedia, "Root-mean-sqaure deviation", 2021 Aug 6th, Accessed on Dec 27th 2021, [Online], Available at: https://en.wikipedia.org/wiki/Root-mean-square_deviation

[9] otexts, "Holt-Winter's seasonal method", Accessed on Dec 27th 2021, [Online], Avialble at: https://otexts.com/fpp2/holt-winters.html

[10] Rob J Hyndman, "Comparing HoltWinters () and ets()", 2011 May 29th, Accessed on Dec 27th 2021, [Online], Available at: https://robjhyndman.com/hyndsight/estimation2/

[11] Free Range Statistics, "Error, rend, seasonality – ets and its forecast model friends", Accessed on Dec 27th 2021, [Online], Available at: http://freerangestats.info/blog/2016/11/27/ets-friends

[12] Machine Learning +, "ARIMA Model – Complete Guide to Time Series Forecasting in Python", 2021 Aug 22nd, Accessed on Dec 27th 2021, [Online], Available at: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

[13] Machine Learning Mastery, "How to Remove Trends and Seasonality with a Difference Transform in Python", 2020 Jun 23rd, Accessed on Dec 27th 2021, [Online], Available at: https://machinelearningmastery.com/remove-trends-seasonality-difference-transform-python/

Part B: Logistics Regression Analysis

Statistics for Data Analytics – Terminal Assignment-Based Assessment

Polina, Prinii
School of Computing
National College Ireland, Mayor Street Lower, IFSC, Dublin 1, D01 Y300, Ireland
x21137757@student.ncirl.ie

Abstract — **This document serves the purpose of a findings report in a business format. The aim is to outline the data, model building steps applied to the final logistics regression and reporting's on the understanding of the relationship between the characteristics and the binary classification.**

Keywords —
SAV File; Multiple Linear Regression; Logistics Regression; Independent Variable; Dependent Variable; SPSS;

Abbreviations and Acronyms —
LR: Logistics Regression
x: Independent Variable
y: Dependent Variable

I. Introduction

The housing market in the United States accounts for over 17% of the United States GDP, raking in a significant amount of $2.8 trillion [1] through the sales of 6.5 million homes in the year of 2020 [2]. The United States has witnessed a steady increase in household sales despite the unpredicted arrival of the global pandemic aka Covid-19, achieving an increase of sales to 7.1 million homes. What's more it is predicted that the United States house market will rise by 6.6% in 2022 [3] when looking at sales, alongside an increase in home prices by an additional 2.9% in conjunction to the increase of 13.2% [4] from 2020 to 2021.

As of 2021, a typical home int the United States averages at $287,148.

Objective

Using Logistics Regression analysis, the report aims to report on the findings derived throughout the steps of:

1. Understanding the variables in the dataset.
2. Model building steps.
3. Rejection of intermediate models.
4. Assumptions have been met and not violated.
5. Final model selection.

With the statistical goal of understanding the relationship between the characteristics and the binary classification as presented within the dataset.

Like all regression analysis, Logistics Regression is a predictive analysis and is appropriate for use when the dependent variable is dichotomous or in other words binary. The Logistics Regression equation is written as follows:

$$\mathrm{E}(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

II. Alternate Approach

The report takes into consideration that the the approach to build a Multiple Linear Regression might assist better when estimating the house characteristics against the binary choice of "Budget" and "Expensive". If a Linear Regression (either on a single or multiple factor) is to be performed on a response of binary nature many complications would arise, them being:

1. The Ordinary Least Squared (OLS) regression would be illogical as the OLS describes the relationships between one or more independent variable against the dependent variable. The OLS is usually denoted by a value between 0 and 1, thus being illogical as a binary response is measure by a 0 and 1.
2. The regression errors could violate Homoscedasticity.
3. Significance assume normal distribution or errors, this cannot be achieved in a binary variable as Y (dependent variable) only has two value, 0 and 1.

Thus, the best fit of regression analysis for the dataset in question is Logistics Regression.

III. Methodology

The analysis aims to perform the following methods to support the understanding and model building of the Logistic Regression with the final step of reporting on the results and findings:

Disclaimer: All works of preliminary assessment, building of the model and selection is undertaken through the software tool SPSS developed by IBM.

A. Logistics Regression Assumptions:

The following assumptions are considered to ensure that the data under analyses using LR is suitable. The assumptions being:

1. The dependent variable is of dichotomous nature (i.e. Sold, Not Sold).
2. Sample size is of suitable size, small samples can be problematic.
3. The data must show absence of Multicollinearity.
4. The data must not contain significant outliers.
5. The data must assume Independence of Errors, that is each response comes from different unrelated case.

Violations in assumptions are resolved as they present, thus the report notes that multiple models may be built in the search of the optimum best fit model.

B. Preliminary Assessment:

Preliminary assessment is undertaken to obtain an understanding of the variables in the HouseCategories.sav dataset.
In conjunction the preliminary assessment evaluates assumptions #1and #2 confirming whether they have been violated or met.

C. Model Building:

The Model Building section looks to cover the methods applied to address any violations of assumptions as well as discuss the steps undertaken to the identification of a suitable Logistics Regression model for the dataset in question.
Assuming assumption #1 and #2 are validated throughout the preliminary analysis, the analysis looks to apply the following methods to address further violations of outstanding assumptions:

- Assumption #3 is met like in MLR analysis, multicollinearity within a LR model can be assessed by the VIF (Variance Inflation Factor) which explains the behaviour (variance) of an independent variable is influenced by other independent variables.
- Assumption #4 is met using Cook's Distance tests for outliers by drafting Box Plots. The said plots identify outliers as ones furthest from the line. Any significant outliers are then removed manually.
- Assumption #5 is met by drafting residuals plots which can confirm that an Independence of Errors is apparent.

The analysis notes that the above steps designated to mitigate assumption violations may be carried out on repetitive accounts. This is to be expected as identifying the optimum model for a Logistics Regression to predict the probability that a given observation falls into either category of a binary y variable based on one or more x variables [5].

IV. Results & evaluations

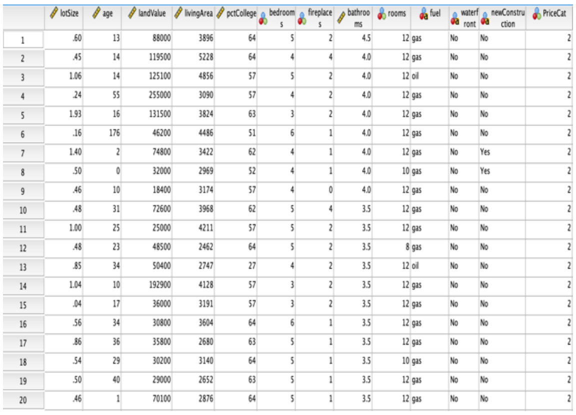
A. Preliminary Assessment Findings:

On the question of preliminary assessment findings have shown that assumptions #1 and #2 have not been violated. Thus, deeming the data is fit for purpose prior to building an LR model, Figure 1 outlines the data under analysis prior to the regression analysis.

Figure 2 provides an outline of the first 20 rows of the data for the readers understanding.

No.	HouseCategories.sav – Data Types			
	Column Name:	Variable Type:	Data Type:	Unit of Measure:
1	lotSize	Independent	Continuous (Scale)	Decimal
2	age	Independent	Continuous (Scale)	Age (0 – 225)
3	landValue	Independent	Continuous (Scale)	Financial
4	livingArea	Independent	Continuous (Scale)	Square Meter
5	pctCollege	Independent	Continuous (Scale)	Percentage
6	bedrooms	Independent	Nominal	Categorical
7	fireplaces	Independent	Nominal	Categorical
8	bathrooms	Independent	Continuous (Scale)	Numeric
9	rooms	Independent	Nominal	Numeric
10	fuel	Independent	Nominal	Categorical
11	waterfront	Independent	Nominal	Binary (Yes/No)
12	newConstruction	Independent	Nominal	Binary (Yes/No)
13	PriceCat	Dependent	Nominal	Binomial (1/2)

Figure 1: HouseCategories.sav – Data Types



	lotSize	age	landValue	livingArea	pctCollege	bedrooms	fireplaces	bathrooms	rooms	fuel	waterfront	newConstruction	PriceCat
1	.60	13	88000	3896	64	5	2	4.5	12	gas	No	No	2
2	.45	14	119500	5228	64	4	4	4.0	12	gas	No	No	2
3	1.06	14	125100	4856	57	5	2	4.0	12	oil	No	No	2
4	.24	55	255000	3090	57	4	2	4.0	12	gas	No	No	2
5	1.93	16	131500	3824	63	3	2	4.0	12	gas	No	No	2
6	.16	176	46200	4486	51	6	1	4.0	12	gas	No	No	2
7	1.40	2	74800	3422	62	4	1	4.0	12	gas	No	Yes	2
8	.50	0	32000	2969	52	4	1	4.0	10	gas	No	Yes	2
9	.46	10	18400	3174	57	4	0	4.0	12	gas	No	No	2
10	.48	31	72600	3968	62	5	4	3.5	12	gas	No	No	2
11	1.00	25	25000	4211	57	5	2	3.5	12	gas	No	No	2
12	.48	23	48500	2462	64	5	2	3.5	8	gas	No	No	2
13	.85	34	50400	2747	27	4	2	3.5	12	oil	No	No	2
14	1.04	10	192900	4128	57	3	2	3.5	12	gas	No	No	2
15	.04	17	36000	3191	57	3	2	3.5	12	gas	No	No	2
16	.56	34	30800	3604	64	6	1	3.5	12	gas	No	No	2
17	.86	36	35800	2680	63	5	1	3.5	12	gas	No	No	2
18	.54	29	30200	3140	64	5	1	3.5	10	gas	No	No	2
19	.50	40	29000	2652	63	5	1	3.5	12	gas	No	No	2
20	.46	1	70100	2876	64	5	1	3.5	12	gas	No	No	2

Figure 2: HouseCategories.sav - First 20 Rows

The above confirms assumption #1 being met, this states that the y variable is of dichotomous nature. In the case of the dataset under analysis the binomial variable is represented as a 1 or 2 with 1 equating to "Budget" and 2 equating to "Expensive".

To confirm on assumption #2, which states that the sample size must be of suitable size. A Bar Chart was plotted against the y variable to facilitate a count of both responses. Figure 3 clearly indicates a suitable sample size as both responses summed reach close to 2000, with a total of 1709 samples available in the dataset.

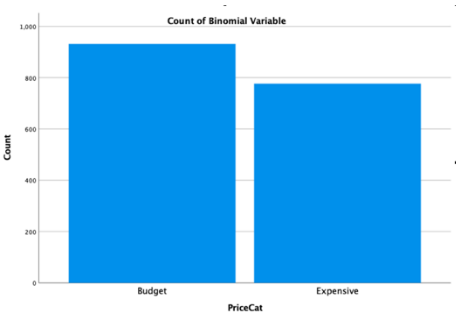
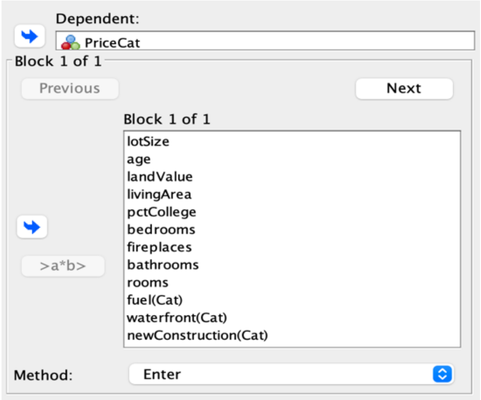


Figure 3: Dependent Variable - Bar Chart

Thus, the preliminary assessment concludes with no violations and both assumptions met.

B. *Model Building Findings:*



Dependent:

PriceCat

Block 1 of 1

Previous | Next

Block 1 of 1

lotSize
age
landValue
livingArea
pctCollege
bedrooms
fireplaces
bathrooms
rooms
fuel(Cat)
waterfront(Cat)
newConstruction(Cat)

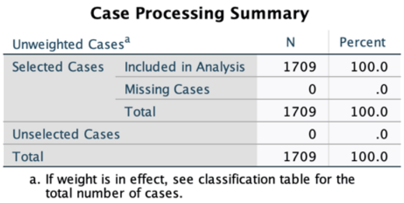
Method: Enter

Figure 4: All x variables against the y variable – Model 1

As iterated previously the model building process takes into consideration multiple attempts in search for the optimum model. The analysis first looked to apply LR analysis to all 12 identified x variables against the y variable as shown by Figure 4.

The model output presented some interesting results in conjunction to the first attempt in search of the optimal model.

For differentiation purposes, the report will denote the first attempt as Model 1.



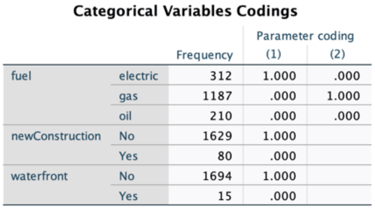
Case Processing Summary

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

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

Figure 5: Model 1 Processing Summary

The descriptive statistics confirmed that Model 1 successfully processed all 1709 records thus, clearing any potential discrepancy issues. Illustrated by Figure 5. Additionally, descritptive statistics allowed for the understanding of categorical variables and the number of occurrence throughout the dataset as illustrated by Figure 6.

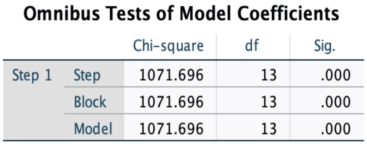


Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
fuel	electric	312	1.000	.000
	gas	1187	.000	1.000
	oil	210	.000	.000
newConstruction	No	1629	1.000	
	Yes	80	.000	
waterfront	No	1694	1.000	
	Yes	15	.000	

Figure 6: Model 1 Categorical Variables

The interpretation of results from Model 1 begins with Block 0, which looks to evaluate the null model or in other words the model with no independent variables in effect. Based on this Model 1 achieved a total of **54.5%** of classification accuracy, a fairly good score however, as independent variables are introduced the analysis looks for this percentage of classification accuracy to increase.



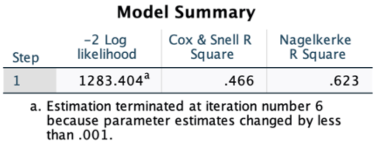
Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1071.696	13	.000
	Block	1071.696	13	.000
	Model	1071.696	13	.000

Figure 7: Model 1 Omnibus Test

The report continues to evaluate the performance of Model 1 with the introduction of x variables in Block 1 however, firstly the report interprets the results of an Omnibus Test of Model Coefficients.

The test provides an overall indication of how well the model inclusive of x variables performs against the null model, the desired result being a Sig. value of less than 0.05. Figure 7 strongly illustrates an excellent goodness of fit as strong significance is present. However, the Hosmer-Lemehow Goodness of Fit Test indicates a poor fit with a significance value of **0.001** which is not above 0.5. To support a model an optimal value grater than 0.5 is desired.



Model Summary

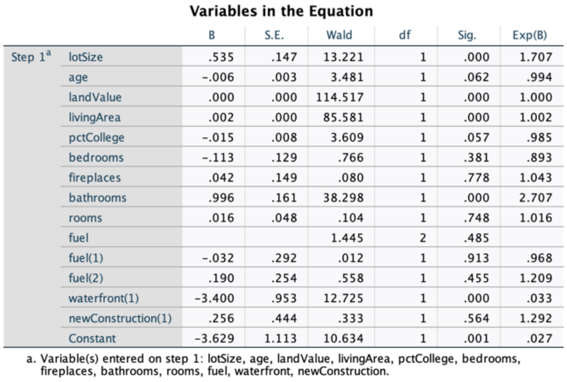
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1283.404[a]	.466	.623

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

Figure 8: Model 1 Summary

This being the first indicator that Model 1 may require some adjustments. This is supported by the Model Summary which outlines the Cox & Snell R Square alongside the Nagekerke R Square. These indicators serve the purpose of explaining the variation in the y variable by the model. Both R Square values are thought of the same as the R Square value from MLR analysis which is popularly used to explain the difference in the y variable cause by the x variables. The results of the R square values as illustrated by Figure 8 are neither strong nor weak. The analysis will continue to identify and build a stronger model.

Strangely enough the classification accuracy improved drastically with the introduction of all 12 x variables. Overall percentage improved to **83.4%**, a strong indicator to the performance of Model 1 in predicting the correct category of "Budget" or "Expensive" for each case.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lotSize	.535	.147	13.221	1	.000	1.707
	age	-.006	.003	3.481	1	.062	.994
	landValue	.000	.000	114.517	1	.000	1.000
	livingArea	.002	.000	85.581	1	.000	1.002
	pctCollege	-.015	.008	3.609	1	.057	.985
	bedrooms	-.113	.129	.766	1	.381	.893
	fireplaces	.042	.149	.080	1	.778	1.043
	bathrooms	.996	.161	38.298	1	.000	2.707
	rooms	.016	.048	.104	1	.748	1.016
	fuel			1.445	2	.485	
	fuel(1)	-.032	.292	.012	1	.913	.968
	fuel(2)	.190	.254	.558	1	.455	1.209
	waterfront(1)	-3.400	.953	12.725	1	.000	.033
	newConstruction(1)	.256	.444	.333	1	.564	1.292
	Constant	-3.629	1.113	10.634	1	.001	.027

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

Figure 9: Model 1 – Variables in the Equation

Lastly, the report looks to evaluate the **Variables in Equation** section as illustrated by Figure 9. Here the report looks for statistically significant variables (<0.05) aka predictors which are assessed using the Wald Statistic (similar to the use of the t-statistic in MLR). Further, the report looks to evaluate the **Exp(B)** values which outines the odds of success within the variable of interest, in this case the report looks to identify which variables hold good odds for a house being classified as "Budget" or "Expensive" based on the characteristics. The values are found in Figure 9.

Assessing the results, a total of 8 characteristic x variables are classified as statistically significant with a p-value less than 0.05, in addition for each significant variable the odds ratio is above 1. Meaning the odds for a house being classified as “Budget” is 1 to 2 times greater than being classified as “Expensive”, this is due to the following characteristics being idenitfied with the support of the Wald Statistic:

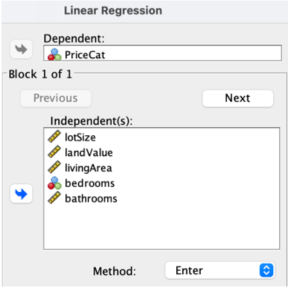
- lotSize
- landValue
- livingArea
- bedrooms
- bathrooms
- fuel(2) aka gas
- waterfront(1), in other words a house includes a waterfront area.
- newConstruction(1)

One statistically insignificant variable – newConstruction(1) with 1 equating to Yes is demonstrating quite interesting results. Though, the variable shows as statistically insignificant, the odds ratio presents itself as above 1. Thus, for these reasons the analysis looks to include the variable in the next iteration of an LR model.

On average Model 1 showed strong evidence of performance for the first attempt, however the Hosmer-Lemehow Goodness of Fit Test idicated a poor fit. The next iteration of the model will look to improve on that evalutation.

The report highlights that no analysis for the verification of assumptions for Model 1 was undertaken, it is assumed that Model 1 was of exploratory nature. Thus, it is believed that the parameters in question regarding model fitness will be improved upon the structure of Model 2.

Prior to proceeding the build of Model 2, the 8 statistically significant x variables within analysis, the report looked to evaluate if assumption #3 has been met against the said x variables. This is achieved through the use of an MLR analysis. The x variables are built against the y variable as illustrated in Figure 10 through MLR, alongside the selection of drafting a collinearity diagnostics which aids in the determination of collinearity between the x variables.


Linear Regression

Dependent: PriceCat

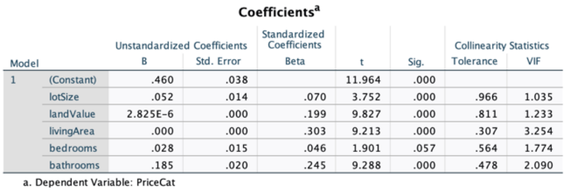
Block 1 of 1

Independent(s):
lotSize
landValue
livingArea
bedrooms
bathrooms

Method: Enter

Figure 10: Model 2 Linear Regression Selection

The desired outcome in this case is a score below 10 demonstarted by the VIF score. The VIF or Variance Inflation Factor explains the behavior (variance) of an independent variable is influenced by other independent variables. Figure 11 illustrates the VIF scores for 5 of 8 x variables in question, the three remaining x variables being: - fuel, - waterfront & - newConstruction cannot be included due to their Categorical nature.


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.460	.038		11.964	.000		
	lotSize	.052	.014	.070	3.752	.000	.966	1.035
	landValue	2.825E-6	.000	.199	9.827	.000	.811	1.233
	livingArea	.000	.000	.303	9.213	.000	.307	3.254
	bedrooms	.028	.015	.046	1.901	.057	.564	1.774
	bathrooms	.185	.020	.245	9.288	.000	.478	2.090

a. Dependent Variable: PriceCat

Figure 11: Model 2 – Multicollinearity Results (Linear Regression)

The VIF values for the Linear Regression with the x variables in interest display a low score thus, ultimately confirming that assumption #3 has not been violated and the data under analysis does not show multicollinearity.

On the topic of assumption clarification the report moves to evaluate assumptions #4 & #5, the results are as follows.

Figure 12 shows that Cook’s Distance identifued a small number of outliers in the data, them being: -949, -1,684 & -553. These are removed manually through the use of SPSS. Having manually removed the outliers, the anlysis found no additional outliers. Thus, the report can confirm that assumption #4 which states that no significant outliers to be present in the dataset has been met.

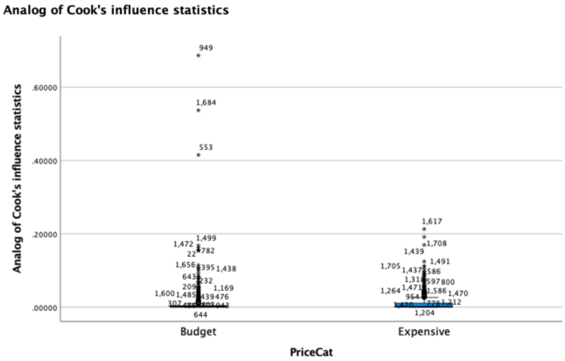


Figure 12: Cook’s Distance – Assumption #4

Lastly, assumption #5 which assumes an Independece of Error is checked by plotting the residuals against a time variables. In the case of this analysis the residuals were plotted against the age of the household [6]. The desired outcome is a random scatter of residuals. Should a non-random patter show then assumption #5 is violated. Figure 13 strongly indicates a random scatter of residuals thus clearing the violation of assumption #5.

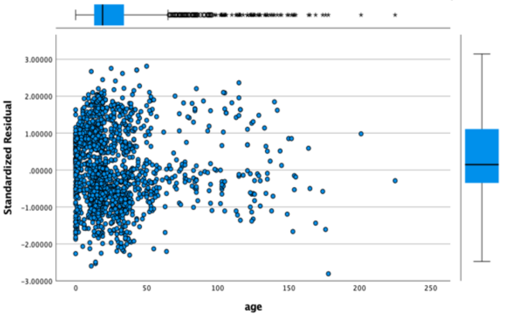


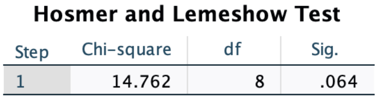
Figure 13: Independence of Errors

Throughout the build of Model 2, a total of 2 assumptions have been violated which have been met and addressed. The report now moves to evaluating the results a Logistics Regression for Model 2.

The report notes that several attempts have been undertaken to idenitfy the optimum fit for the Logistic Regression. The driving factor behind the idication of fit for Model two was the Hosmer and Lemeshow Test, with the end goal of producing a p-value greater than 0.05. This was achieved after 3 attempts which resulted in the exclusion of the categorical type variables being:

- fuel
- waterfront
- newConstruction

With the introduction of Model 2, the analysis was successful in achieving a **.064** p-value. A significant increase from .001 from Model 1. Figure 14 illustrates the results for Hosmer and Lemeshow Test for Model 2. Though, the report takes into consideration that numerous problems are present with the test, such as the lack of consideration for overfitting [7]. However, a non significant value was achieved thus, indicating fitness of model as good being the interest of the analysis.

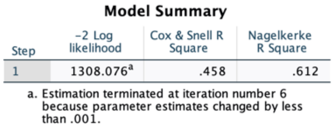


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.762	8	.064

Figure 14: Model 2 Hosmer & Lemeshow

A minor change to the Omnibus Test of Model Coefficients occurred with the introduction of Model 2, having retained its significance factor the Chi-Square dropped from 1071.696 to **1047.024** alongside the degrees of freedom dropping from 13 to **6**.



Model Summary

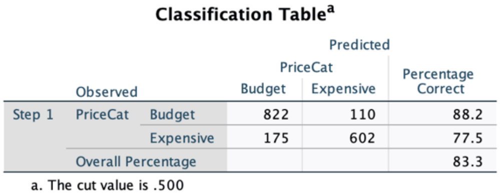
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1308.076a	.458	.612

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

Figure 15: Model 2 Summary

The Model Summary for Model 2 also showed improvement, though the pseudo-R square values remained relatively similar the log likelihood increased. This is a good indication as the log likelihood supports the goodness of fit for a given model [8]. The higher the value the better a model fits its given dataset, in this case the HouseCategories.sav dataset. Alongside the Hosmer and Lemeshow Test, the report can conclude that an overall good fit has been achieved with Model 2 as illustrated by Figure 15.

To support and conclude on Model 2 findings the report evaluates the Classification Table. With the introduction of Model 2 the classification accuracy reduced by **0.1%**, an extremely little reduction leaving a fairly small impact on the performance of Model 2. Figure 16 illustrates the classification accuracy of Model 2.



Classification Tablea

	Observed		Predicted		
			PriceCat		Percentage Correct
			Budget	Expensive	
Step 1	PriceCat	Budget	822	110	88.2
		Expensive	175	602	77.5
	Overall Percentage				83.3

a. The cut value is .500

Figure 16: Model 2 – Classification Table

The results in this chapter indicate that the optimum model has been identified being Model 2 thus, concluding the search for the best fit model. In addition, all relevant assumptions have either been met or addressed if violated. Throughout the chapter the significant result have been highlighted supporting the identification of Model 2.

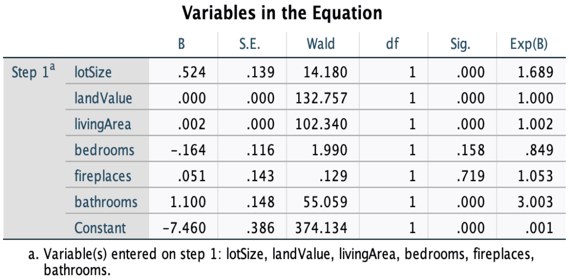
The next chapter, therefore, moves to conclude and report on the findings of Model 2.

V. Conclusion

In conclusion a Logistic Regression was performed with the intention to assess the impact of several factors to determine on the likelihood that a household would be categorized as “Budget” or “Expensive”. In the end the model contained a total of five independent variables (lotSize, landValue, livingArea, bedrooms and bathrooms). The full model reported that all entered predictors were statistically significant reporting that x^2 (6, N = 1709) = 1047.024, $\rho < .000$, a strong indication that the built model was successful in distinguishing between house classifications of “Budget” or “Expensive”.

As a whole, the model was able to explain between 46% (Cox and Snell R square) and 61% (Nagelkerke R Suqare) of the variance of “Budget” or “Expensive” and correctly classified 83% of household classifications. As illustrated in Figure 17, all variables for the exception of one (fireplaces) contributed statistical significance to the model. The strongest predictor being bathrooms, reocrding an odds ratio of 3.1 which indicated that a household was 3 times more likely to be classified as “Budget” with less bathrooms in the household.

This is supported by the preivous Figure 3 from Preliminary Assessment which clearly showed that houses are more greatly classified as “Budget”.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1a	lotSize	.524	.139	14.180	1	.000	1.689
	landValue	.000	.000	132.757	1	.000	1.000
	livingArea	.002	.000	102.340	1	.000	1.002
	bedrooms	-.164	.116	1.990	1	.158	.849
	fireplaces	.051	.143	.129	1	.719	1.053
	bathrooms	1.100	.148	55.059	1	.000	3.003
	Constant	-7.460	.386	374.134	1	.000	.001

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

Figure 17: Model 2 – Variables Summary

VI. Acknowledgements

I would like to express my appreciation to National College of Ireland who has provided me the possibility to complete this report. A special thanks I give to our lecturer Hicham Rifai who teaches Statistics for Data Analytics, as well as Tony Delaney for producing the course material online. Both have helped me to coordinate my project, especially in understanding the objective and in writing this report.

VII. References

[1] Congressional Research Service, “Introduction to U.S. Economy: Housing Market”, 2021 May 3rd, Accessed on Dec 28th 2021, [Online], Available at: https://sgp.fas.org/crs/misc/IF11327.pdf

[2] Statista, “ Total home sales in the U.S. 2011-2022”, 2021 May 4th, Accessed on Dec 28th 2021, [Online], Available at: https://www.statista.com/statistics/275156/total-home-sales-in-the-united-states-from-2009/

[3] Forbes, “Experts Predict What The Housing Market Will Look Like in 2022”, 2021 Dec 13th, Accessed on Dec 28th, [Online], Available at: https://www.forbes.com/sites/brendarichardson/2021/12/13/experts-predict-what-the-housing-market-will-look-like-in-2022/?sh=3886ca3f3942

[4] CNBC, “The typical home price is up a record 13.2% compared to last year, according to Zillow”, 2021 Jun 16th, Accessed on Dec 28th, [Online], Available at: https://www.cnbc.com/2021/06/16/typical-us-home-price-up-record-13point2percent-compared-to-last-year.html

[5] aerd statistics, “Binomial Logistics Regression using SPSS Statitics”, Accessed on Dec 28th, [Online], Available at: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

[6] uTexas, “Using Plots to Check Model Assumptions”, Accessed on Dec 30th, [Online], Available at: https://web.ma.utexas.edu/users/mks/statmistakes/modelcheckingplots.html

[7] Statistics How To, “Hosmer-Lemeshow Test: Definition”, 2016 Aug 28th, Accessed on Dec 30th, [Online], Available at:https://www.statisticshowto.com/hosmer-lemeshow-test/

[8] Statology, “How to Interpret Log-Likelihood Values (With Examples)”, 2021 Aug 21st, Accessed on Dec 30th, [Online], Available at: https://www.statology.org/interpret-log-likelihood/