Forecasting of Harmonised Consumer Index of Prices Using Time Series Forecasting Models

Heena Chopra
MSc in Data Analytics
National College of Ireland
Student No. : x19205309

***Abstract*—This report describes a Time Series model for predicting the Harmonised Index of Consumer Price using Time series models i.e., ARIMA Model, Holt's Winter's Seasonal Model and STL (Seasonal and Trend decomposition using Loess) Model based on its previous values listed in form of a dataset provided by the European Union (EU) dataset Repository. The Data Cleaning, Statistical Analysis and Model Development for the report was done using R.**

I. INTRODUCTION

Harmonised Index of Consumer Prices (HICP) measures the changes in price of household goods and services over time. These are basically used to compare the change (increase/decrease) in prices of products acquired by households taking into account the harmonised definitions. HICP indicates the changes in prices of goods and services and their stability for ECB (European Central Bank). HICP is basically the CIP i.e., Consumer Price Index which is compiled together according to methodologies defined across different countries. Forecasting HICP is a crucial aspect if one wants to formulate better decisions for their organisation. We will be using various time series forecasting models for accomplishing this task.

II. DATA SOURCES, CLEANING AND PREPARATION

A. Data Source

- **Variables Used:** Harmonic Index of Consumer Price (HICP).
- **Data Source:** The Dataset for HICP was taken from European union data repository in .csv format.

B. Cleaning and Preparation

The data was first read from a csv file and saved into an appropriate file object. After that, the column heads were renamed accordingly. A time series object was then constructed from the data using the ts () function. The start and end values for the dates were provided in the argument and the frequency was specified as 12(since the data is monthly). The time series object was then plotted using the autoplot () function for further analysis. The data contains 264 observations ranging from January 1998 to December 2019.

III. MODEL THEORY

Time Series is a set of values of a particular variable measured at equal intervals of time (For E.g., Monthly, Quarterly, Yearly). Using Time Series Forecasting, we can analyse any time-based series of data to forecast its future values. A Time Series has a number of features such as White Noise, Seasonality, Trends, Stationarity which need to be analysed before modelling of the data.

Auto Correlation is one of the major aspects that comes into play while analysing a time series. ACF (Auto correlation function) and PACF (Partial Auto Correlation Function) plots are used for this purpose. These two plots together provide the complete information about the correlation between different values of a time series. The ACF plot provides the information about how the present values in a time series are correlated with values in the past. The PACF plot finds the correlation between the residuals which are left after all the patterns have been explained by the earlier lags.

Stationarity is the most important aspect for Time Series analysis. A Time Series is said to be stationary if it has a constant mean and variance and covariance is independent of time. If the statistical properties of a series are not changing with time then it is said to be stationary in nature. Dicky Fuller Augmented test is run to check if a series is stationary or not. If the result of the Test shows the series to be stationary (i.e., p¡0.05) then we proceed with the further analysis. If the test results show the series to be non-stationary (i.e., p¿0.05) then it can be made stationary using differencing. A series which gets stationary after differentiating it once is said to be integrated of order 1 and is denoted by I(1). Similarly, if the series gets differentiated twice to make it stationary it is denoted by I(2) and so on. So, in general a stationary series that is differentiated d number of times is denoted by I(d) and is called to be integrated of order d.

After analysing all the necessities for a Time Series Modelling, we apply the following models to time Series Data.

1) **ARIMA Model**: ARIMA stands for Auto Regressive Integrated Moving Average. It a time series model that forecasts the future values on the basis of its own lags (Past Values) and the lagged errors. An ARIMA model comprises of 3 values (p, q, d):
   - p is the order of auto regressive term.
   - q is the order of moving average term.

- d is the integrated order required to make the series stationary.

.

2) **Holt's Winter's Seasonal Model**: This method was invented by Holt (1957) and Winters (1960) to capture seasonality in the data. This method comprises of 3 forecast and smoothing equations each for seasonality, trend and levelling along with some smoothing parameters.
3) **STL Model**: STL stands for seasonal and trend decomposition using Loess. STL model works on the principle of decomposing a time series data into 3 components that contain seasonality, trend and residual. Locally estimated scatterplot smoothing (LOESS) is used by STL method to uproot all the smooth estimates from the components of the model.

The accuracy for these models is examined using 2 tests:

1) **Normal Q-Q Plot**: This is to test the Normality in the Data.
2) **Ljung-Box Test**: This test is applied on the residuals after fitting the model to determine whether or not there is any remaining structure in the residuals.

IV. Model Analysis

Time Series forecast models were coded in R. The steps for same are stated as below:

A. Decomposition and Smoothing

1) After Cleaning and preparation of the data, decomposition of time series was done in order to see the different elements that make up the series such as seasonality of the series, trend of the series and the white noise pervading the time series. Ref to figure 1.
2) Stationarity of the time series was tested using Augmented Dickey-Fuller Test which resulted in p-value to be 0.01 and stated the series to be stationary integrated with the order 0 i.e.; I (0). Ref to figure 2.
3) ACF and PACF plots for the time series is shown in figure 3 and 4 respectively.
4) Ref to figure 5 that is depicting two series of data where one represents the original data and other one shows the moving average of the data, done to make the data series smoother.

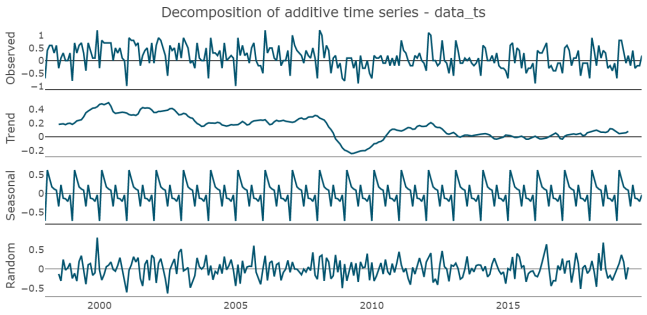

Fig. 1. Additive decomposition for time series

```
Warning message in adf.test(data_ts, k = 2):
"p-value smaller than printed p-value"

        Augmented Dickey-Fuller Test

data:  data_ts
Dickey-Fuller = -9.1868, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

Fig. 2. Stationarity(ADF) test for time series

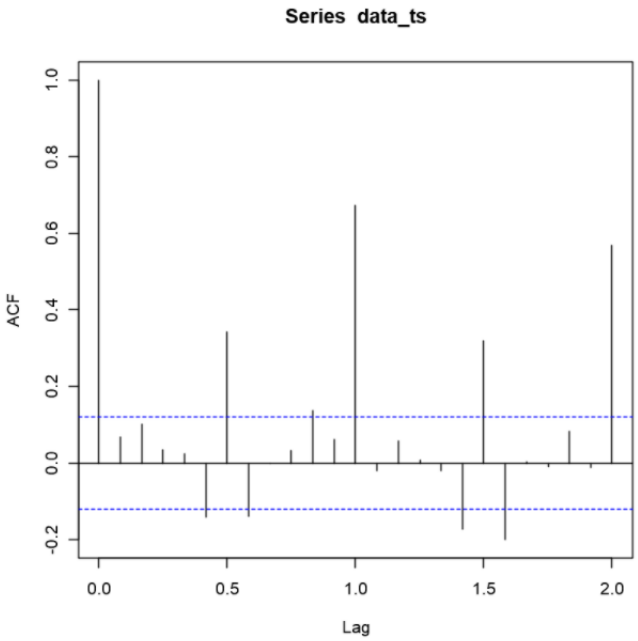

Fig. 3. ACF plot for time series

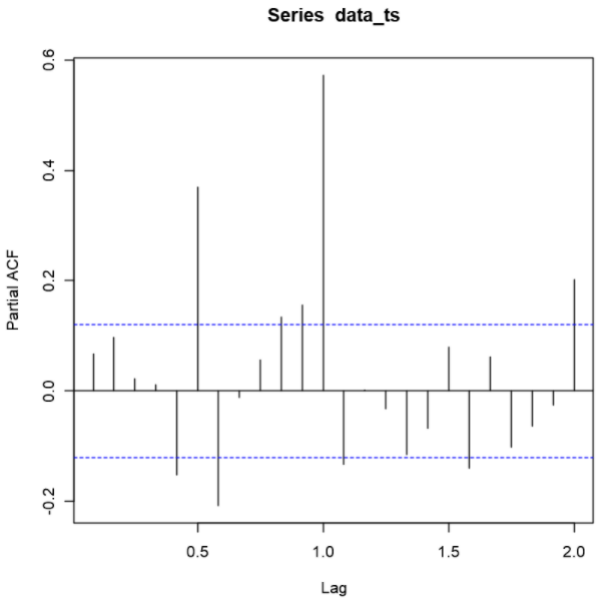

Fig. 4. PACF plot for time series

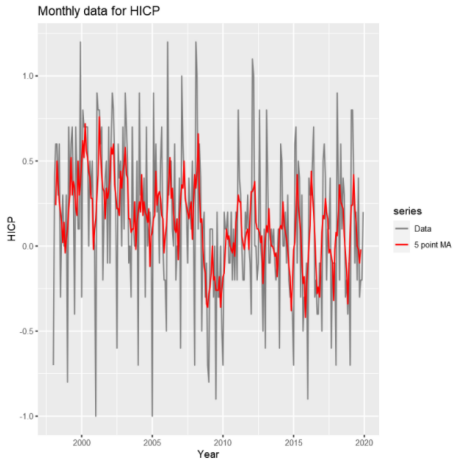

Fig. 5. 5 point moving average plot for time series

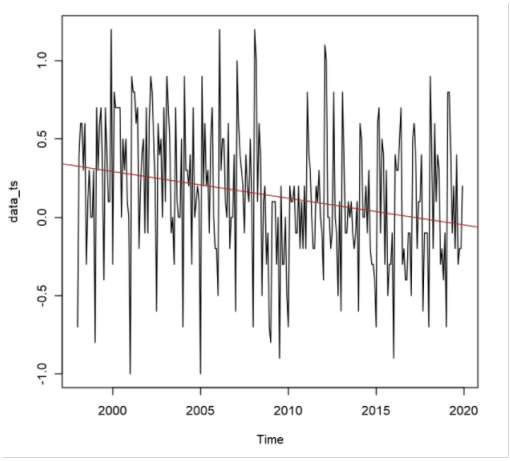

Fig. 6. Linear fitted line for time series

B. Modelling

1) ARIMA Model: The Auto ARIMA function suggested ARIMA(0,0,1)(2,1,2)[12] with drift to be the best model.(refer Fig 7)

```
 Best model: ARIMA(0,0,1)(2,1,2)[12] with drift

Series: data_ts
ARIMA(0,0,1)(2,1,2)[12] with drift

Coefficients:
         ma1     sar1     sar2     sma1     sma2    drift
      0.1839  -0.1371  -0.0472  -0.5775  -0.2000  -0.0013
s.e.  0.0604   2.0275   0.3103   2.0300   1.7515   0.0004

sigma^2 estimated as 0.0767:  log likelihood=-37.29
AIC=88.59   AICc=89.05   BIC=113.29
```

Fig. 7. ARIMA(0,0,1)(2,1,2) model parameters

The residuals follow the normal distribution as depicted by the Fig 8.

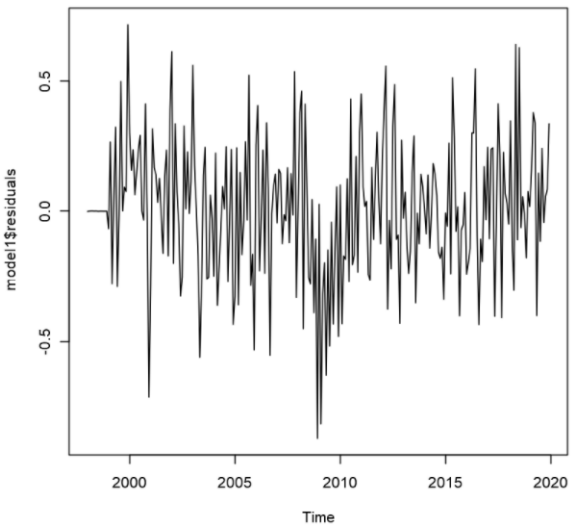

Fig. 8. Plot of residuals vs time for ARIMA(0,0,1)(2,1,2) model

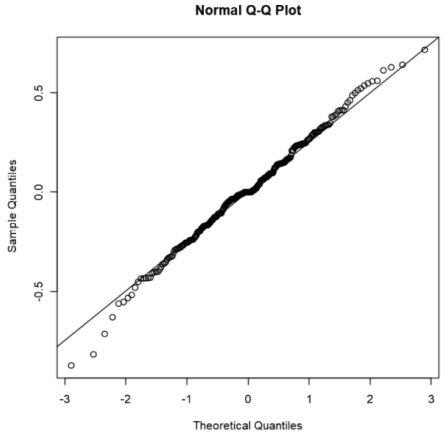

Fig. 9. Normal Q-Q plot for ARIMA(0,0,1)(2,1,2) model

```
        Ljung-Box test

data:  Residuals from ARIMA(0,0,1)(2,1,2)[12] with drift
Q* = 31.254, df = 18, p-value = 0.0269

Model df: 6.   Total lags used: 24
```

Fig. 10. Ljung-Box test for ARIMA(0,0,1)(2,1,2) model

A matrix: 1 × 7 of type dbl

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.006961134	0.2673451	0.2063518	NaN	Inf	0.8062119	0.009388624

Fig. 11. ARIMA(0,0,1)(2,1,2) model accuracy parameters

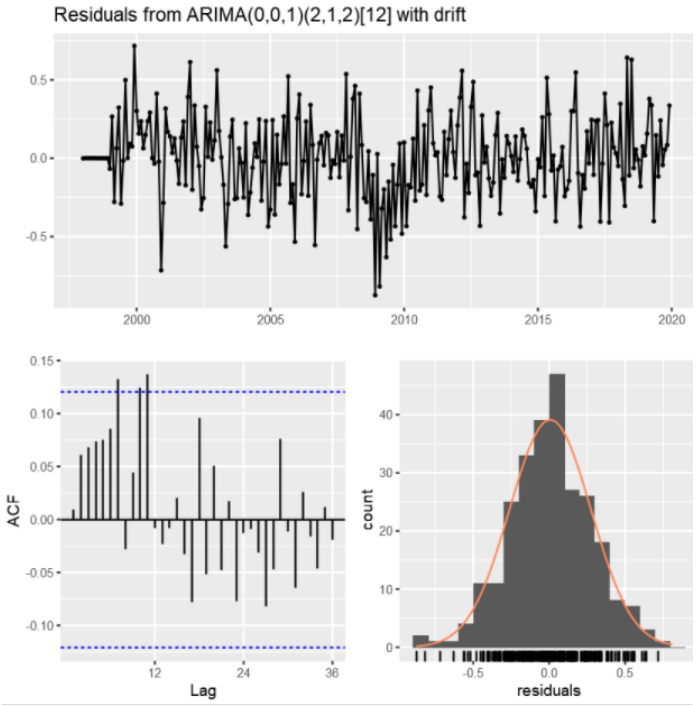

Fig. 12. Residual plots for ARIMA(0,0,1)(2,1,2) model

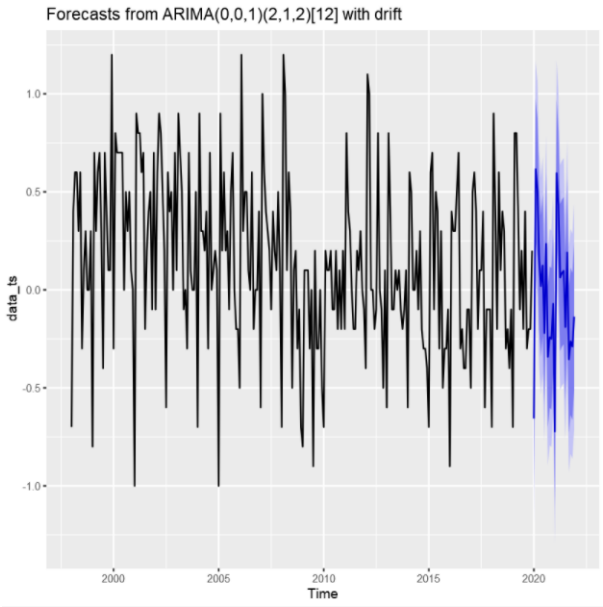

Fig. 13. Forecasts from ARIMA(0,0,1)(2,1,2) model

2) Holt's Winter Seasonal Model: The Holt's winter seasonal method is usually used to model the seasonality, trend and average components of a time series. It can be applied to both additive and multiplicative decomposed series. Figure 14 below shows the forecasting obtained from the model.

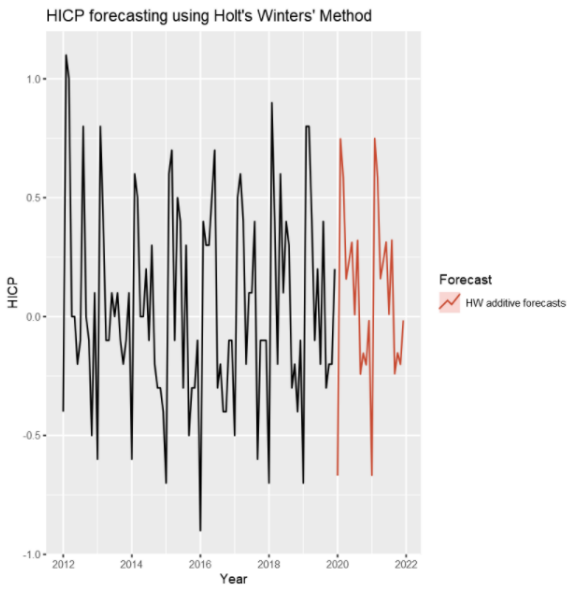

Fig. 14. Holt's Winter Forecasting

3) STL Decomposition Forecasting Model: STL uses the LOESS(Local Regression Method) to model the Seasonality and Trend components. Here, we have first decomposed the series using STL decomposition(Figure 16 and then forecasted the value for for year 2020.

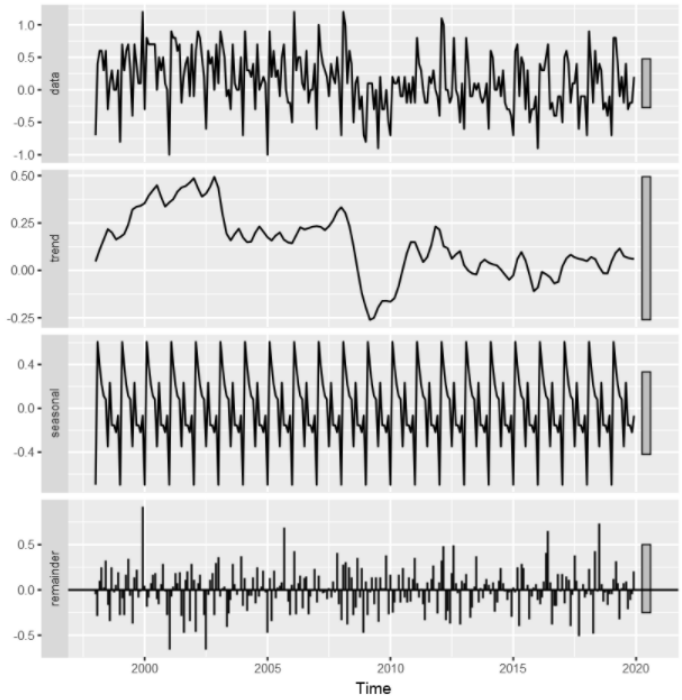

Fig. 15. STL Decomposition of the time series

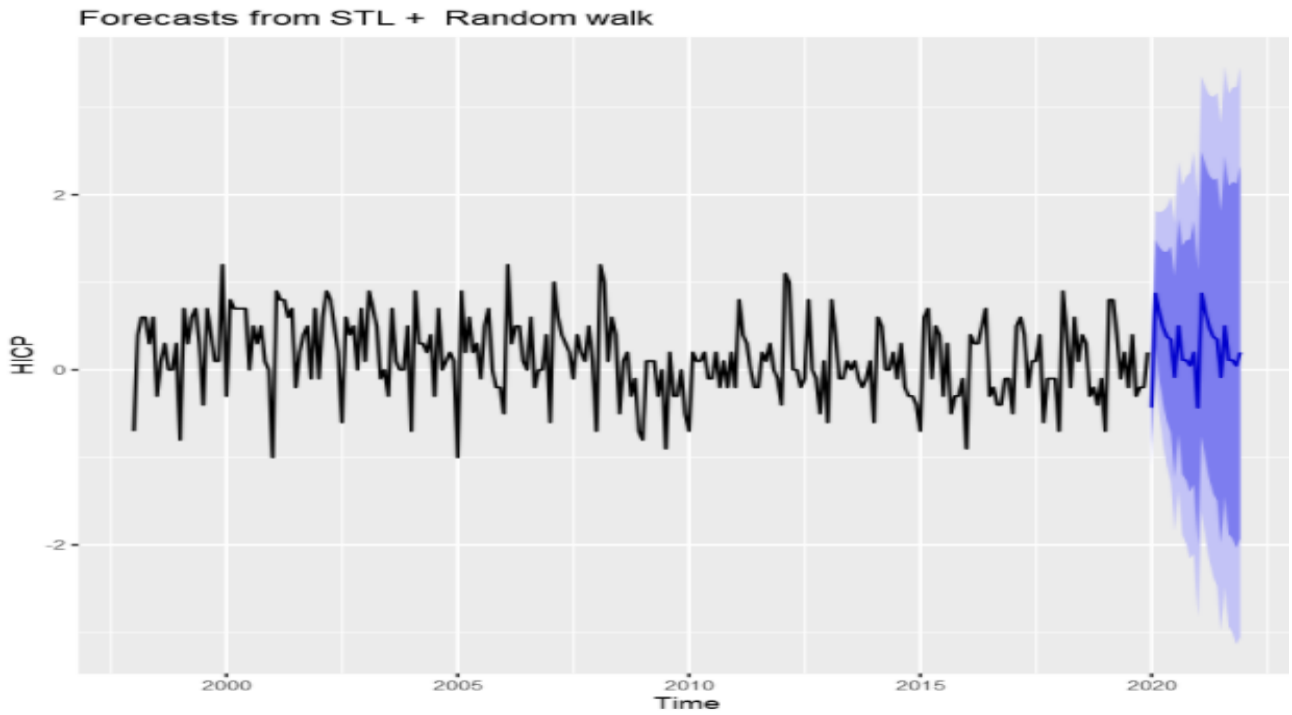

Fig. 16. Forecast from STL Decomposition

Prediction of School Readiness of Young Children Using Binary Logistic Regression

***Abstract*—This part of the project describes a binary logistic regression model which is used for predicting the school readiness of young children. In other words, we aim to predict if a child can read most or all of the alphabets using various variables like their age, their TV viewing hours, etc. The model is evaluated using various parameters like *Wald's Test*, *Cox Snell pseudo R squared* and *Hosmer and Lemeshow test*. The analysis was done using IBM SPSS software and the cleaning of data was performed using python.**

I. INTRODUCTION

The early years in a child's age contribute considerably to their development as a whole. These early years are very important in a child's overall health and development. The overall development of a child not only includes the social, emotional and health aspects but also their educational requirements as well. Therefore it is very crucial for a child to be taught things appropriate for the scope of their age. Teaching a child alphabets and numbers would naturally be the first step for anyone undertaking the responsibility of their all round development. Therefore it becomes extremely crucial to analyze the various factors upon which a child's ability to read the alphabets could depend on. These include(but not limited to) their TV viewing habits, whether their mother or father read to them and their age. We will built a model based on these variables to accurately predict whether a child would be able to read most or all of the alphabets(1) or not(0). The model will then be analysed and evaluated by various parameters.

II. DATA SOURCE, CLEANING AND PREPARATION

A. Variables used and Data Source

1) *Data Source:* The dataset for logistic regression was compiled from various surveys conducted and published by Pew Research Centre on their webpage.
2) *Variables Description:*

Variables	Data Type	Description
age	Numeric	Child's Age
TVHOURS	Numeric	Hour's child watches video/tv
age3	Numeric	Child is 3 years Old
age4	Numeric	Child is 4 years Old
age5_6	Numeric	Child is 5 years Old
count20	Numeric	Counts to 20 or more
letters	Numeric	Knows most/All Letters
momread	Numeric	Read by Mom
dadread	Numeric	Read by Dad

Fig. 1. Description of variables used

3) *Cleaning and Preparation:* The dataset for logistic regression was extracted from the pew website and was cleaned and pre-processed using NumPy and Pandas libraries in Python. The independent variables were renamed accordingly. The null values were removed from the dataset and the final dataset was imported into a .csv file.

III. MODEL THEORY

Logistic Regression is a regression analysis which is applied on a dataset when we have a dependent variable that is dichotomous in nature and a number of independent variables that can be ordinal, nominal, interval and ratio levelled in nature. This is the go-to model for classification problems where we have two classes of data. The general equation for Logistic regression is:

$$P = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_n X_n)}} = \frac{1}{1 + e^{-(\beta_0+\sum \beta_i X_i)}}$$

Logistic Regression allows evaluation of multiple variables by using the basic principles. This method is similar to linear regression where we have a response variable and many predictor variables. Unlike Linear regression, where the response variable can either be continuous or categorical but in logistic regression the response variable has to be categorical (Binary). Logistic Regression model has an S-shape illustration i.e., it has a logistic curve which is limited to have values either 0 or 1.

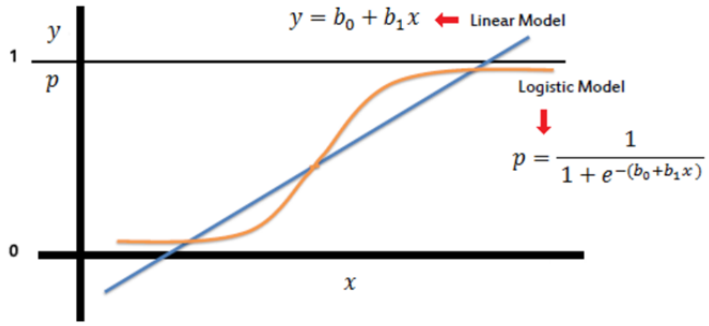

Fig. 2. Logistic and Linear Model

IV. MODEL ANALYSIS

A. Model 1

1) **Dependent Variables**

- Letters

2) **Independent Variables**
   - TV Hours
   - Age
   - Age3
   - Age4
   - Age56
   - Count20
   - Momread
   - Dadread

3) **Model Statistics**
   - *Pseudo R square*:
     - Cox and Snell R square= 0.242
     - Nagelkerke R square=0.330
   - *Correlation Values*: All the values are less than 0.8, Ref to Figure 3.
   - *Coefficient p-values*:
     - Age =0.999
     - Age3= 0.999
     - Age 4=0.999
     - Momread =0.247

     All other coefficient p values are less than 0.05.
   - *Model tests* :
     - p- value (Wald's Test) < 0.05
     - p-value (Hosmer and Lemeshow Test) < 0.05
     - p-value (Chi-Square Test) < 0.05
   - *Accuracy* = 75.9 %

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.533	.045	141.657	1	.000	1.704

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	591.855	7	.000
	Block	591.855	7	.000
	Model	591.855	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2226.699[a]	.242	.330

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

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	24.628	8	.002

Fig. 3. Model 1:Tests Summary

B. Model 2(Final Model

1) **Dependent Variables**

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	age	19.174	14120.893	.000	1	.999	212486597.0	.000	.
	TVHOURS	-.012	.005	6.389	1	.011	.988	.979	.997
	age3	36.997	28241.785	.000	1	.999	1.169E+16	.000	.
	age4	18.617	14120.893	.000	1	.999	121698568.4	.000	.
	count20	1.897	.110	295.165	1	.000	6.667	5.369	8.278
	momread	.203	.175	1.342	1	.247	1.225	.869	1.727
	dadread	.491	.108	20.789	1	.000	1.634	1.323	2.017
	Constant	-96.027	70604.463	.000	1	.999	.000		

a. Variable(s) entered on step 1: age, TVHOURS, age3, age4, count20, momread, dadread.

Classification Table[a]

			Predicted letters 0	1	Percentage Correct
Step 1	Observed letters	0	465	326	58.8
		1	190	1158	85.9
	Overall Percentage				75.9

a. The cut value is .500

Correlation Matrix

		Constant	age	TVHOURS	age3	age4	count20	momread	dadread
Step 1	Constant	1.000	-1.000	.000	-1.000	-1.000	.000	.000	.000
	age	-1.000	1.000	.000	1.000	1.000	.000	.000	.000
	TVHOURS	.000	.000	1.000	.000	.000	.046	.019	.076
	age3	-1.000	1.000	.000	1.000	1.000	.000	.000	.000
	age4	-1.000	1.000	.000	1.000	1.000	.000	.000	.000
	count20	.000	.000	.046	.000	.000	1.000	-.072	-.078
	momread	.000	.000	.019	.000	.000	-.072	1.000	.028
	dadread	.000	.000	.076	.000	.000	-.078	.028	1.000

Fig. 4. Model 1:Full Summary

   - Letters $\hat{Y}$

2) **Independent Variables**
   - Age (x_1)
   - TV Hours (x_2)
   - Count20 (x_3)
   - Dadread (x_4)

3) **Model Statistics**
   - *Model Equation*:

$$\hat{Y} = \frac{e^{-3.487+0.726(x_1)-0.012(x_2)+1.909(x_3)+0.487(x_4)}}{1+e^{-3.487+0.726(x_1)-0.012(x_2)+1.909(x_3)+0.487(x_4)}}$$

   - *Pseudo R square*:
     - Cox and Snell R square= 0.241
     - Nagelkerke R square=0.329
   - *Correlation Values*: All the values are less than 0.8, Ref to Figure 6
   - *Coefficient p-values*:

     All other coefficient p values are less than 0.05.
   - *Model tests* :
     - p- value (Wald's Test) < 0.05
     - p-value (Hosmer and Lemeshow Test) < 0.05
     - p-value (Chi-Square Test) < 0.05
   - *Accuracy* = 75.6 %

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.533	.045	141.657	1	.000	1.704

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	588.650	4	.000
	Block	588.650	4	.000
	Model	588.650	4	.000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.725	8	.678

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2229.905[a]	.241	.329

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Fig. 5. Model 2 : Tests Summary

Classification Table[a]

			Predicted letters		
	Observed		0	1	Percentage Correct
Step 1	letters	0	457	334	57.8
		1	188	1160	86.1
	Overall Percentage				75.6

a. The cut value is .500

Correlation Matrix

		Constant	age	TVHOURS	count20	dadread
Step 1	Constant	1.000	-.909	-.203	-.115	-.276
	age	-.909	1.000	-.053	-.112	.078
	TVHOURS	-.203	-.053	1.000	.048	.077
	count20	-.115	-.112	.048	1.000	-.076
	dadread	-.276	.078	.077	-.076	1.000

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	age	.726	.082	79.359	1	.000	2.067	1.762	2.426
	TVHOURS	-.012	.005	6.393	1	.011	.988	.979	.997
	count20	1.909	.110	300.811	1	.000	6.747	5.438	8.372
	dadread	.487	.107	20.498	1	.000	1.627	1.318	2.008
	Constant	-3.487	.317	121.106	1	.000	.031		

a. Variable(s) entered on step 1: age, TVHOURS, count20, dadread.

Fig. 6. Model 2: Full Summary

Principal Component Analysis

I. INTRODUCTION

Large datasets are increasing gradually with a widespread increase in many other disciplines as well. Such large datasets have to be interpreted in a way that the dimensionality of the dataset gets reduced without hampering most of the information from the data. We have a number of methods and techniques for this purpose but Principal Component Analysis is one of the oldest and widely used technique. Principal Component Analysis is a method for reducing the dimensionality of the data but at the same time preserving most of the variability of the data.

Principal Component Analysis finds new variables that are the linear functions of the values present in the dataset. These variables are uncorrelated with each other and successfully maximize the variance. These new variables constructed are called as the PC Components and are made in such a way that most of the information of the data is squeezed into the first components, keeping the maximum values in the first component and then second and so on, like shown in the picture below.

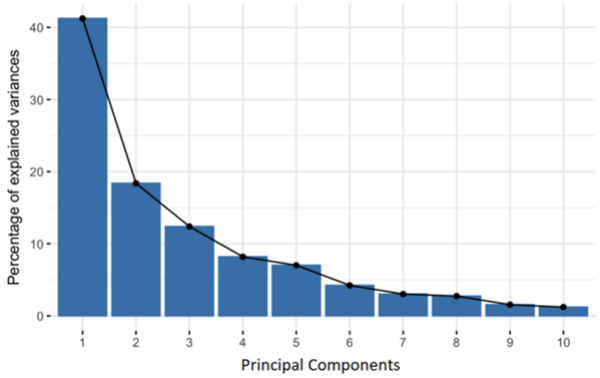

Fig. 1. Illustration of PCA

The main task of Principal Component Analysis is identification of the patterns in the data and highlighting its similarities and differences to give it a direction. It makes the data easy to explore, analyse and visualise. This report further explains Principal Component Analysis by using a real time Dataset and its analysis using PCA.

II. DATA SOURCE, CLEANING AND PREPARATION

A. Source of Data and Variables description

- **Data Source:** PCA was demonstrated using Wine Dataset that was taken from UCI Machine Learning Repository in .csv format. This Dataset contains the results of chemical analysis done on the wines grown in same areas of Italy but by 3 different cultivators. The analysis established the proportion of all the 13 constituents present in all the three types of wine.
- **Variables Used:** The Dataset contains 14 variables where Wine Class is the Class Identifier i.e., three types of wines (1-3) and rest all variables demonstrates the amount of its proportion used in making the wines.

Variables	Data Type
WineClass	int
Alcohol	num
Malic acid	num
Ash	num
Alcalinity of ash	num
Magnesium	int
Total phenols	num
Flavanoids	num
Nonflavanoid phenols	num
Proanthocyanins	num
Color intensity	num
Hue	num
OD280/OD315 of diluted wines	num
Proline	int

Fig. 2. Variables Description

B. Cleaning and Preparation

The data was first read from a csv file and saved into an appropriate file object. After that, the column heads were renamed accordingly. A Principal Component object was then constructed from the data using the prcomp () function. The start and end values for the variables were provided in the argument.

C. Analysis

After Cleaning and Preparation of the Dataset, The Data was first plotted pair wise to see the correlations between all the variables. The graph plotted also shows the pairwise interactions of all the variables with each other.

As we can clearly see in the below graph, it's hard enough to look at all the pairwise visualisations personally and analyse the information from the data. For Multi-Dimensional data like this there is no other meaningful way to analyse the data. In these type of cases PCA is of great help.

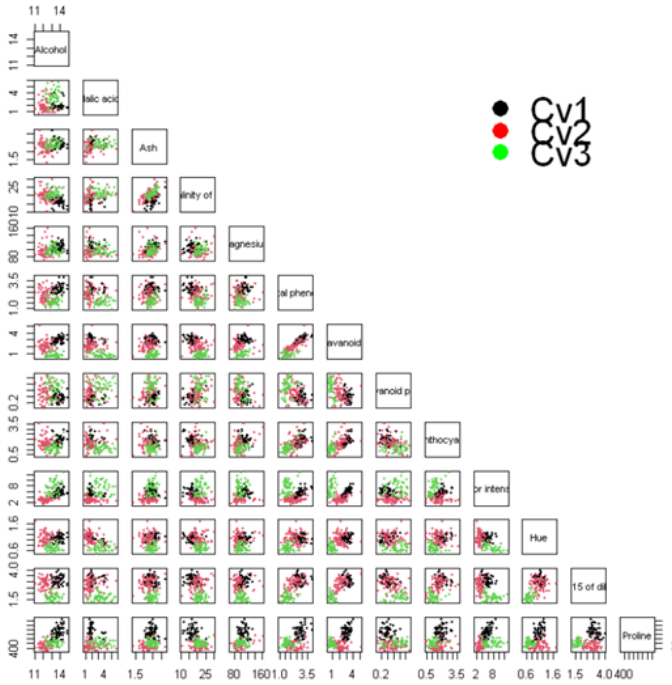

Fig. 3. Pairwise interactions of variables with each other

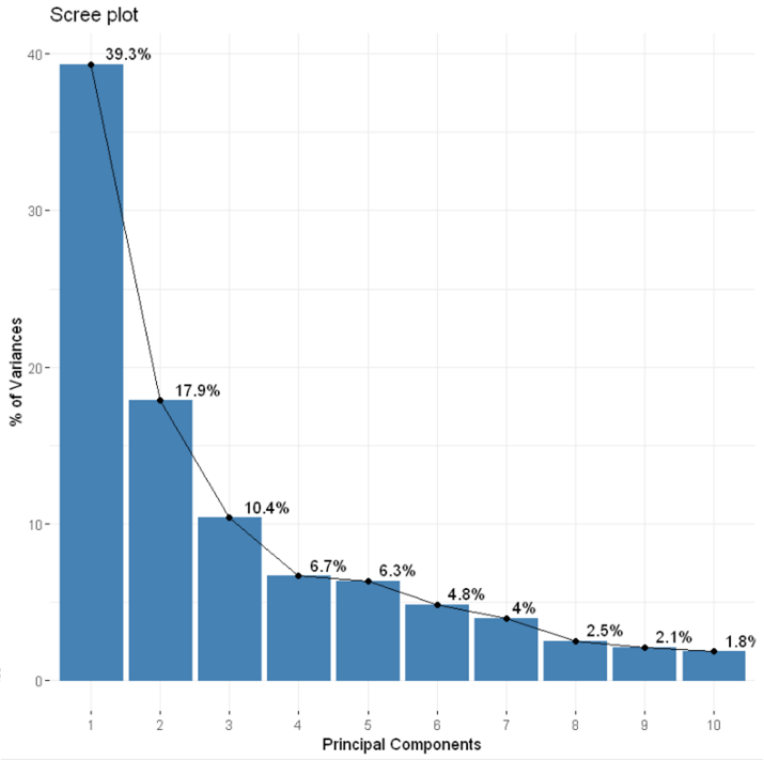

Fig. 5. Scree plot of principal components for the data

Now, PCA is applied to the dataset resulting into 14 PCA's where we can clearly see that the first few Principal Components (PC1 – PC3) contain the most of the proportion of variance.

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.3467 1.5816 1.2055 0.96554 0.94207 0.82191 0.74512
Proportion of Variance 0.3933 0.1787 0.1038 0.06659 0.06339 0.04825 0.03966
Cumulative Proportion  0.3933 0.5720 0.6758 0.74242 0.80582 0.85407 0.89373
                          PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.59350 0.54377 0.50767 0.47523 0.41034 0.36051 0.24077
Proportion of Variance 0.02516 0.02112 0.01841 0.01613 0.01203 0.00928 0.00414
Cumulative Proportion  0.91889 0.94001 0.95842 0.97455 0.98658 0.99586 1.00000
```

Fig. 4. Principal Component Analysis output in R

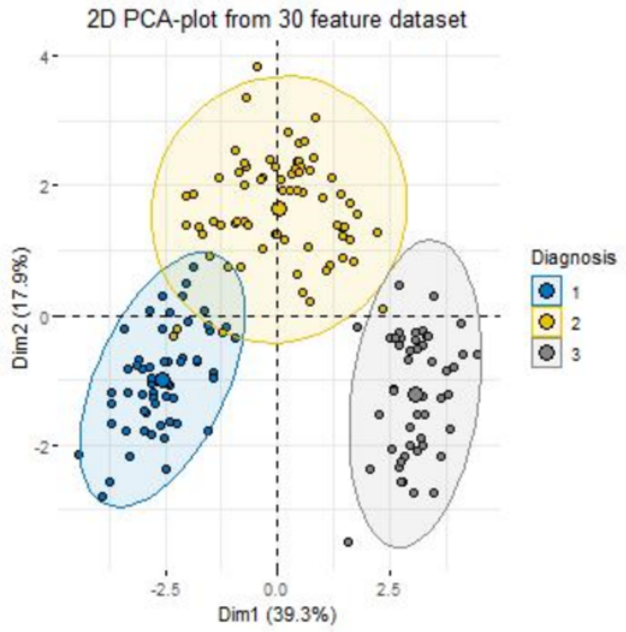

Fig. 6. Cluster Plot for PC1 and PC2

Figure 5 explains the percentage of variances for each of the principal components. It shows which principal component holds how much of variance. This would be helpful in analysing how much of the variance each PCA holds out of the total variance of the dataset.

Figure 6 explains how just 2 principal components are good enough to explain the separation between the three Wine Classes. This plot depicts that we have captured 57.29 % (Dim1-17.9% and Dim2-39.3%) of the variance of the entire dataset using just 2 Principal Components i.e., PC1 and PC2.

The figure 7 shows that first 6 Principal Components have explained 90% of the variance of the data. This means that we can successfully reduce the dimensionality (Number of Variables) from 13 to 6 while losing only 10% of the total variance.

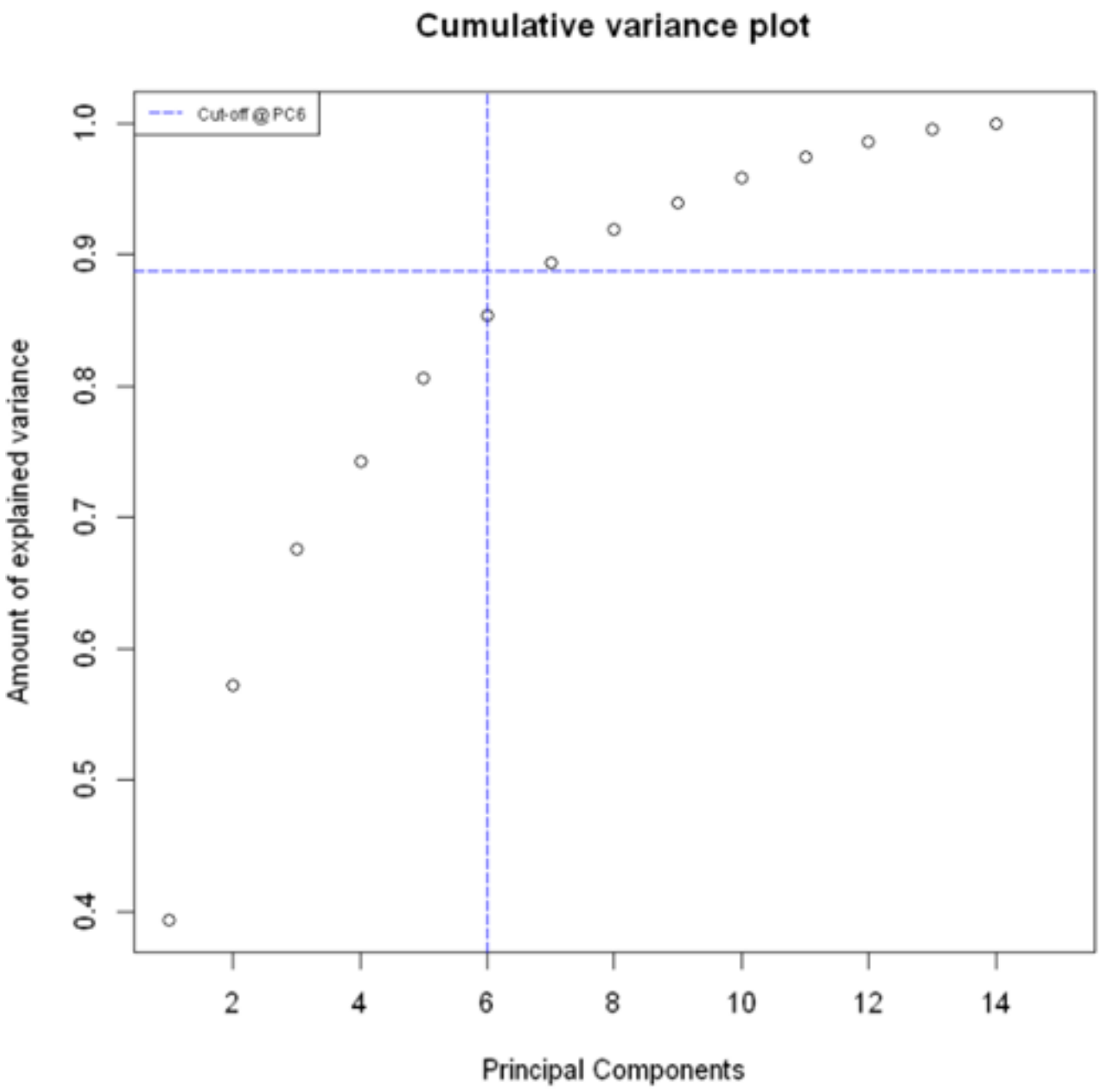

Fig. 7. Cumalative Variance Plot