Study on Time-Series Analysis, Logistic Regression and Principal Compononent Analysis using Statistical Tools and Concepts

1st Abhishek Sunil Padalkar
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19221576@student.ncirl.ie

***Abstract*—A complete statistical analysis is performed for time series analysis on a time series data, logistic regression model building on a binary classification data, and principal component analysis on a complex correlated dataset. In all these analysis, diagnostic checks are performed to then do evaluation and interpretation.**

***Index Terms*—time series analysis, logistic regression, binary classification, principal component analysis, diagnostics**

I. Introduction

This project work consists of three unrelated sections of statistical analysis. The first section focuses on Time Series Analysis on an import trade dataset for Ireland. In the second section, a generalised logistic regression model is built and evaluated with interpretation of the variables. In the final section, for a highly complex Superconductor dataset with 41 variables, a dimension reduction using Principal Component Analysis is performed and interpreted. All these individual analysis is taken based on statistical concepts and tools to achieve a good acceptable result.

II. Time Series Analysis

This section of this statistical analysis report focus on time series analysis. The aim for this analysis is, on a selected data, to create 3 time series candidate models, perform required checks for evaluation of each model and compare all models based on RMSE and AICc accuracy metrics for better fitting model. After comparison, select an optimal model based on in-sample forecast accuracy and perform forecast for 3 time periods ahead using the selected model. To perform this analysis, the platform used is R.

A. Data Description

The dataset used for the purpose of time series analysis is Ireland import trade in Euros every year from 1988-2019 (31 periods). Trade value is in millions of EUR.

B. Basic level understanding of Time Series data

Before building the models, it is important to understand the time series data as there can be components such as seasonality, trends, cycles or irregular fluctuations present in it. By performing these checks, it becomes easier to prioritize different types of models that can fit the given data well. First, the given time series is simply plotted. Fig. 1 shows the graphical true representation of the given time series data.

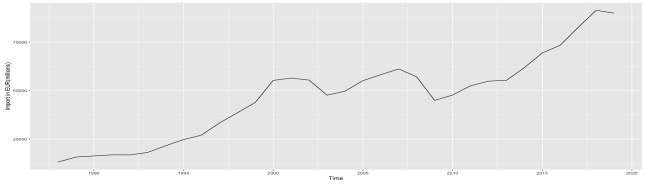


Fig. 1: Time Series plot of Ireland Import Trade

It shows a clear pattern of an upward trend in our data, thus the time series is not stationary. Since the time series is yearly based, it is expected that the series will not have seasonality in it. From the fig. 1, as the time series is already interpreted as having some trend in it, we don't need to smooth it using moving averages technique.

C. Tests to check any components in our Time Series

1) Check for Trend: Augmented Dickey-Fuller(ADF) test is performed to check if the time series is stationary or not. For a significant result implies that the time series is stationary or else has trend in it. Fig. 3 shows that our time series is not stationary and is confirmed that it has an upward trend.

2) Check for Seasonality: With an intuition that our yearly data probably will not have seasonality in it, a check is performed for confirmation. Simply using seasonal plot function in R for decomposition, it is confirmed that there is no seasonality in the data. Fig. 4, shows an error in decomposing the time series.

3) Check for Cycles: Cycles are components in time series which are different than seasonality as the cycle length is not constant like seasonal components. These components can be seen by looking at any cyclic pattern with respective to trend line. It is checked by using Hodrick-Prescott Filter on our time series data. Fig. 5 demonstrates that there is a cyclical component deviating from the trend line. As confirmed above, our time series has a trend and cyclical pattern in it and no seasonality component. Thus, to include trend in our forecast, we fit a Holts model, an ETS model, and an ARIMA model to our time series. A naïve model is also fitted to check how well the above sophisticated model performs.

```
> adf.test(irl_import_yearly)  # p-value = 0.53 > 0.05 => Our data has trend

        Augmented Dickey-Fuller Test

data:  irl_import_yearly
Dickey-Fuller = -2.1006, Lag order = 3, p-value = 0.534
alternative hypothesis: stationary
```

Fig. 3: Augmented Dickey Fuller Test on Time Series

```
> seasonplot(irl_import_yearly) # => Error: Data are not seasonal
Error in seasonplot(irl_import_yearly) : Data are not seasonal
```

Fig. 4: Failure in Seasonal Decomposition of Time Series

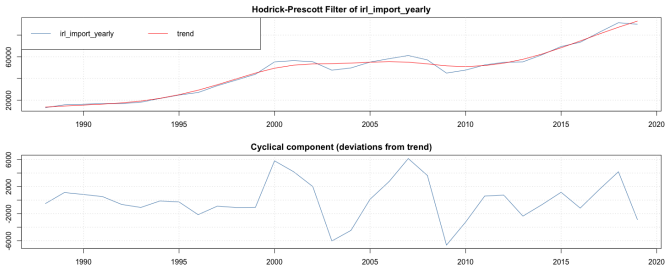


Fig. 5: Hodrick-Prescott Filter Test on Time Series

D. Modeling the Time series

Here, we fit our time series in 4 different models, a Naïve model, Holts Model, ETS model and an ARIMA Model.

1) Naïve Model: Naïve forecasting method is a simplistic method where, a forecast made for a period is same as the previous period observed in the series. Fig. 2a shows the summary of the model. Naïve model's in-sample error measure RMSE is 5257.606.

2) Holts Model: Holts model is a type on exponential smoothing(ES) model. These models assigns exponentially diminishing weights on a the previous observations such that total of all weights is 1. It takes a value of alpha as weight for the previous period to the current time period, and then assigns diminishing weights according to each periods accordingly. Alpha value needs to be selected so that correct amount of weightage is given to the recent previous observation and thus other previous observations accordingly. In Holts model, a beta weight is also considered taking trend into account for a given time series. Similar to alpha, it's necessary to select a best beta value for better fitting model. Using holt function in R, these alpha and beta values selected are the best values possible for the fit. Fig. 2b, shows the summary of our Holts model. The best alpha and beta values for respective weights are 0.9999 and 10^-4. This tells us that almost all, 0.99, of the weightage is given to the previous time-series period for the forecast. AICc metric is 662.50 and in-sample RMSE is 4566.371 for this model.

3) ETS Model: ETS model is a type of model which can fit the time series based on Errors, Trend and Seasonality. The function in R takes 3 arguments for each of the above 3 factors. We can pass either N, A, M, or Z to each of the 3 arguments. N implies not to include the respective factor, A and M implies whether the respective factor is Additive or Multiplicative in nature, and Z can be put to check all the combinations of models and to select the best model. For this analysis purpose, as our data is not seasonal we put N in the Seasonal factor argument and Z in both Errors and Trend factor to get the best subset model. Fig. 2c shows the summary of the ETS model. The best selected model suggest that, Errors considered are Multiplicative in nature and the Trend factor is Additive in nature. Again in this model, like Holts model, alpha and beta value selected is 0.9999 and 10^-4 respectively. The AICc measure is 648.97 and in-sample RMSE is 4574.832.

4) Arima Model: Arima Models are sophisticated time series models which takes autoregressive, moving averages and differencing terms into account while fitting a time series data. Autoregressive(AR) elements are the previous time series periods with certain weights assigned to each lag. This is different than ES model as here we need not select all the periods observed in the past for our forecast with diminishing weights but a few periods back of the current forecast and assign weights to each period to get a best forecast fit accordingly. In Arima, Moving Averages(MA) elements are the error terms for previous period forecasts. By including previous errors made takes into account for any recent shock in the series, thus helps to make a better forecast. Number of lagged terms and or previous error terms needed to be include in a Arima model is denoted by p and q respectively. Arima models can only be correctly fitted to a time series data if it's stationary in nature. Thus, for any time series having trend in it needs to difference it as required number of time to fit it

```
> # Naive Model
> naive_yearly <- naive(irl_import_yearly, h=1)
> summary(naive_yearly)

Forecast method: Naive method

Model Information:
Call: naive(y = irl_import_yearly, h = 1)

Residual sd: 5257.6056

Error measures:
                    ME     RMSE      MAE      MPE
Training set 2481.529 5257.606 4187.277 5.58323
```

(a) Naive Model

```
> # Holts Model
> holt_yearly <- holt(irl_import_yearly, h=1)
> summary(holt_yearly)

Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = irl_import_yearly, h = 1)

  Smoothing parameters:
    alpha = 0.9999
    beta  = 1e-04

  Initial states:
    l = 9704.6463
    b = 2406.8973

  sigma:  4881.656

     AIC     AICc      BIC
660.1979 662.5056 667.5266

Error measures:
                    ME     RMSE      MAE       MPE     MAPE
Training set 103.1062 4566.371 3187.207 -0.6998464 7.131897
```

(b) Holts Model

```
> # ETS Model
> ets_yearly <- ets(irl_import_yearly, model = "ZZN")
> summary(ets_yearly)
ETS(M,A,N)

Call:
 ets(y = irl_import_yearly, model = "ZZN")

  Smoothing parameters:
    alpha = 0.9999
    beta  = 1e-04

  Initial states:
    l = 9705.332
    b = 2212.7734

  sigma:  0.0963

     AIC     AICc      BIC
646.6694 648.9771 653.9980

Training set error measures:
                    ME     RMSE      MAE       MPE
Training set 296.9268 4574.832 3229.134 -0.1426523 7.2
```

(c) ETS Model

Fig. 2: Model Summaries

using ARIMA model. Number of times differencing needed is denoted by d. These, p,d,q, are the arguments an ARIMA model takes to fit a time series data. ARIMA models are thus sophisticated to build and difficult to interpret. For our time series data, we check for number of differencing needed to make our data stationary. Fig. 7, shows that only 1 differencing is needed to make the series stationary.

```
> paste("No. of differencing needed: ", ndiffs(irl_import_yearly))
[1] "No. of differencing needed:  1"
```

Fig. 7: No. of differencing to make Time Series stationary

Series is seen as fluctuating at a fairly constant mean and is horizontal. To check if our time series is stationary we perform ADF Test. Fig. 8 shows that according to the ADF test our data is still not stationary. Another differencing is performed

```
> adf.test(irl_import_yearly_diff) # p = 0.3587 => Indicates that there is still a trend in the data

        Augmented Dickey-Fuller Test

data:  irl_import_yearly_diff
Dickey-Fuller = -2.5559, Lag order = 3, p-value = 0.3587
alternative hypothesis: stationary
```

Fig. 8: Augmented Dickey Fuller Test on Differenced Time Series

on the current differenced data which gives the ADF p-value as significant, 0.04019.
Thus, value for d can be either 1 or 2 based on our two results above. Thus, we check models for both the d values. To select values for p and q, ACF and PACF plots of the differenced series are referred. ACF and PACF are the autocorrelation and partial-autocorrelation plots of Lags in the time series. Fig. 6 shows ACF and PACF plots for 1 order and 2 order differenced time series data respectively.

It can be seen that, for 1 order differenced series, in both the ACF and PACF plots, there is no gradual decrease of significant spikes as well as no significant spikes at the first few lags. This is misleading to understand which values of p and q are suitable for this series data. But, there is a spike at lag 1 in both the plots which is very close to the significant dashed line. Thus, we use d=1 and 3 combinations of p and q, to have p,d,q as 111, 110, and 011. And, for 2 order differenced series, again we see no significant spikes or any gradual decrease in the significant spikes in both the plots. But, there is spike at lag 2 which is almost reaching the significant line. Thus, for d=2 we use 3 combination of p and q, to have p,d,q as 222, 220, 022. We also use Auto Arima function to then find the best fitting model as we also not want to violate principal of parsimony by including many AR and MA terms unnecessarily. Fig. 9 shows the combined performance of all 7 arima models fitted, where red line is the AIC fit metric and blue line is RMSE performance measure.

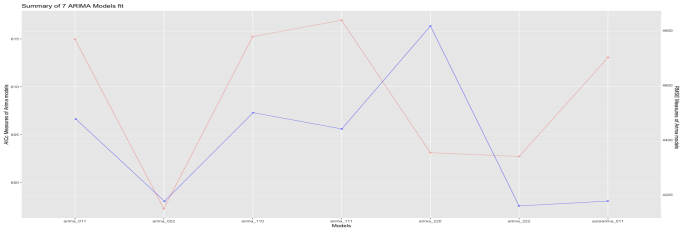
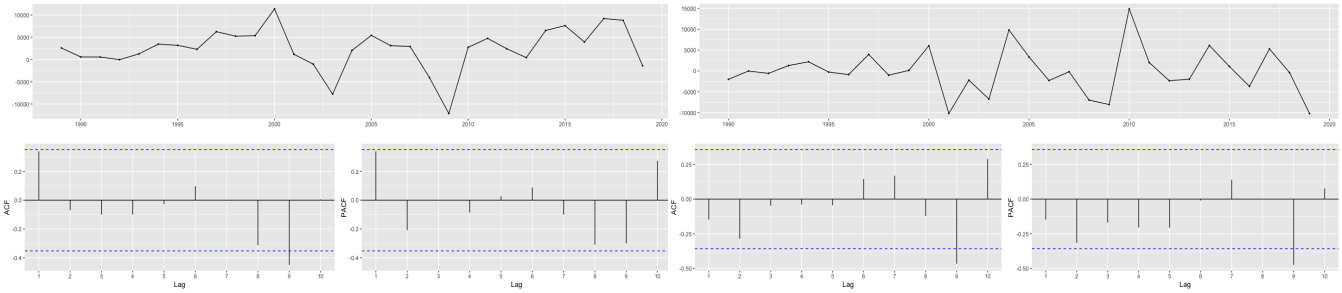


Fig. 9: Summary of all 7 ARIMA Model Performance,

We can see that RMSE for ARIMA models 022, 222 and autoarima 011 drift are low and the AICc metrics for models 022, 220, and 222 are low. Thus, we can see that higher values for p and q with 2nd order differencing fits the data well based on AICc. Arima Model 022 has the lowest AICc value of 597.26 and second lowest RMSE of 4182.855. Auto Arima model has a drift element added to it giving AICc of 613.07 and RMSE of 4183.375. It shows that for 1 order differencing, using only 1 MA term with drift element can give a good fit. We can see that RMSE values of Arima 022 and Auto Arima are very close to each other but AICc values have a good amount of difference. Thus, for a better fitted model, we select Arima 022 as our final ARIMA model.

E. Comparison Between the Models

In this section, the Naïve, Holt, ETS, ARIMA, fitted models are compared based on AICc metric of better fit and RMSE of forecast accuracy measure. Referring back to fig. 2 and fig. 9, we see the AICc and RMSE values for our 4 models. The ARIMA model 022 has the lowest AICc of 597.26 as compared to Holts and ETS models. RMSE measures is again lowest for ARIMA 022 model of 4182.855. Where as Naïve model has RMSE of 5257.6 which is the highest. So, our sophisticated ARIMA model is doing far better than just a naïve model. Holts and ETS models have AICc of 662.5 and



(a) Differenced once

(b) Differenced twice

Fig. 6: ACF and PACF plots for differenced Time Series

648.97 respectively with around RMSE values closer to 4570. Thus, based on these two measures and model performance, ARIMA 022 is selected as our optimal best fitted model.

F. Evaluating ARIMA022 Model

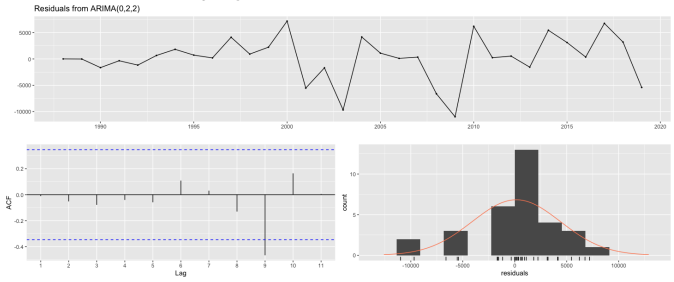
For time series models to be generalized, we have to perform certain checks on the residual errors of the model. There are 2 criteria to be followed by a time series model based on its residuals which are: 1. Residual errors must be normally distributed at the mean value of 0 and must have a constant standard deviation. 2. There should not be any autocorrelations between the residual errors of each lag. If a model passes these two criteria then the model can be said to be an appropriate time series model. To check if there is any autocorrelation between the the residual lags, we perform Ljung-Box test. If this test is significant then there are autocorrelations between the residual errors and if not then otherwise. Fig. 10a and 10b show the Ljung-Box test and the Residual distribution and ACF plot for the ARIMA 022 model respectively.

```
        Ljung-Box test

data:  Residuals from ARIMA(0,2,2)
Q* = 1.0266, df = 4, p-value = 0.9057

Model df: 2.   Total lags used: 6
```

(a) Ljung-Box test of Autocorrelation



(b) ACF and Residual plot of ARIMA 022

Fig. 10: Model Evaluation

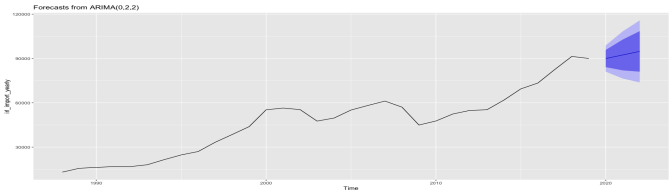
The Ljung-Box test p-value for our model is 0.9057, thus our residuals are simply white noise with no correlation between them. Also, our residuals show a reasonable normal distribution with mean around zero and constant standard deviation. Thus, this final model is said to be an appropriately fitted time series model to the given data. A note is made here, in Fig. 10b we can see there is a spike at lag 9 of residual ACF plot which is a little beyond the significant dashed line. This indicates there is some chance that this model maybe have missed any patterns in the series. That can be the cyclic pattern discussed in the earlier subsection. But, all the other lags have no significant spikes, which may also indicate that the spike at lag 9 is still a white noise. Being cognizant of this fact we consider this model as our appropriate generalized model based on the evaluation performed.

G. Forecasting using ARIMA022

Using our final evaluated ARIMA model, 3 forecasts ahead of the current period are performed. Fig. 11a and 11b shows the 3 point forecast with 80% and 95% confidence intervals range of values and Forecast plot using ARIMA 022 model.

```
> print(fcast_arima)
     Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
2020       89986.70 84166.15  95807.25 81084.94  98888.46
2021       92401.38 81962.83 102839.94 76437.00 108365.77
2022       94816.07 81087.82 108544.31 73820.53 115811.61
```

(a) Forecast values



(b) Model Forecast with CI of 80% and 95%

Fig. 11: Forecast made for 3 years ahead using ARIMA 022

The 3 forecast values for 2020, 2021, 2022 are 89986.7, 92401.38, 94816.07 EUR in millions respectively of import trade for Ireland.

III. Logistic Regression

This project part focuses at creating an appropriate generalized logistic regression model for classifying if a person is satisfied or not with the way things are happening in his/her country. The aim of this work is to create a general good fitted logistic regression model following all the assumptions of Maximum Likelihood Estimator method, evaluation and interpretation of the final model for overall population based on the given sample data. To achieve that with the given data, a dependent variable is selected so that it is mutually exclusive and exhaustive, a logistic regression model is finalized based on variable significance and fitted to given data. This work is performed entirely in R using R studio.

The topic of interest chosen to perform this task is Individual person satisfaction with his or her country based on how things are currently going on. It is interesting to know that given a set of variables whether the person is satisfied or dissatisfied with it's own country. But, that is not the primary aim rather we want to fit a model on this "satisfaction" data and perform checks which are discussed below.

A. Objectives of the analysis

1) Select the sample data from Pew Research website and apppropriate dependent and independent variables.
2) Perform descriptive statistics on the data before modelling.
3) Build a logistic regression model on the data and remove the insignificant variables if any for finalizing the model based on Wald's Statistic.
4) Perform diagnostic check on the final model.
5) Evaluate the model based on comparison with null and predicted accuracy, Hosmer and Lemeshow goodness of fit test, Sensitivity and Specificity, Residual Deviance and Pseudo R^2 measures.
6) Summarise and Interpret the final model and explanation of Odds Ratio for all the variables.

B. Data Description

As discussed above, with the interest in predicting a person's satisfaction with its country, the sample data gathered from the Pew Research Center website is a survey data of Global Attitudes in Spring 2019. As the main criteria for binary logistic regression is that the dependent variable must be binary, mutually exclusive and exhaustive, the question in the survey whether a person is satisfied with its country is chosen as the dependent variable. The other independent variables, survey questions, are chosen by simple intuition that the variable might be significant relating to our dependent question. This project is performed being cognizant of the fact that there can be other more significant variables which can also help in building a better model than with just current variables.

In the Survey Data downloaded, for cases where an individual did not answered or didn't know the answer to any questions were removed. This data was then sampled down from 22000 observations to 250 observations for this project purpose.

1) Description of variables:

- COUNTRY_SATIS: This is our dependent variable having two values, "Satisfied" and "Dissatisfied" which are encoded as 0 and 1. Thus, our model will predict probabilities for a person to be dissatisfied or not.
- ECON_SIT: This survey question checks how good or bad does the individual thinks the economy situation is of his or her country. It is an ordinal variable with categories as "Very good", "Somewhat good", "Somewhat bad", and "Very bad".
- CHILDREN_BETTEROFF2: This question asks if an individual thinks that the children in the same country will be better or worse financially in the future. It's a binary variable with category "Better off" and "Worse off".
- SATISFIED_DEMOCRACY: Asks the individual how satisfied he or she is with the way democracy is working in their country. Values in this variable ranges from "Very satisfied", "Somewhat satisfied", "Not too satisfied", "Not at all satisfied".
- FUTURE_EDUCATION: This question asks, looking at the future of the country, does the individual feel optimistic or pessimistic about the education system of its country. It's a binary variable holding two values "Optimistic" and "Pessimistic".
- MARKET_ECON: How far an individual agrees to say that most people do very good in a free market even though some people are rich and some are poor. The values ranges from "Completely agree", "Mostly agree", "Mostly disagree", "Completely disagree".
- FREE_ELECTIONS: Asks the individual that how important he or she thinks is to have a honest elections held with a choice of at least 2 parties in their country. The variable holds "Very important", "Somewhat important", "Not too important", "Not important at all".
- SEX: This variable describes gender of the individual where "Male" is 1 and "Female" is 2

As seen from above, except for "CHILDREN_BETTEROFF2", "FUTURE_EDUCATION", and "SEX", all the other independent variables are ordinal in nature.

C. Descriptive Statistics

Before model building, a descriptive statistics is performed for understanding the data better. Fig. 12 shows a description of the satisfaction data. The dataset has 250 observations and a total of 7 independent variables and 1 dependent variable, "COUNTRY_SATIS". The mean of 0.56 of COUNTRY_SATIS indicates that there are more individuals who are dissatisfied with their country than those who are satisfied. For ordinal variables discussed above, min is 1 and max value is 4. ECON_SIT has mean of 2.52, thus there are very less individuals who think their country's economic condition is "Very good". Most of think that it is "Somewhat good". SATISFIED_DEMOCRACY with mean of 2.59 indicates that there are very less individuals who think are "Very satisfied" with their countries democracy. Many are "Somewhat satisfied" or not. MARKET_ECON mean is 2.26 indicating many "Mostly agree" that most of the population are better off with the free market economy. FREE_ELECTIONS mean is 1.4 conveying that most people think it is "Very important" to have an honest elections held in their country. For binary variables, min is 1 and max is 2. CHILDREN_BETTEROFF2 mean is 1.51 which says that there's a almost 50% of individuals who thinks children are "Better off" in the future. FUTURE_EDUCATION has a mean of 1.43 indicating more people are "Optimistic" considering education in the future in their country. With 1.43 as mean for SEX, there are somewhat more men than women who took the survey. By considering all this information, a simple idea is formed that individuals who are dissatisfied with the current country situation can be male who are optimistic about the future education but they are somewhat satisfied with the country's economic condition and democracy.

```
> ### Descriptive Stats ###
> describe(ga_sampled_250)
                    vars   n mean   sd median trimmed  mad min max range
ECON_SIT               1 250 2.52 0.88      2    2.52 1.48   1   4     3
CHILDREN_BETTEROFF2    2 250 1.51 0.50      2    1.51 0.00   1   2     1
SATISFIED_DEMOCRACY    3 250 2.59 0.90      3    2.62 1.48   1   4     3
FUTURE_EDUCATION       4 250 1.43 0.50      1    1.42 0.00   1   2     1
MARKET_ECON            5 250 2.26 0.96      2    2.21 1.48   1   4     3
FREE_ELECTIONS         6 250 1.40 0.73      1    1.23 0.00   1   4     3
SEX                    7 250 1.43 0.50      1    1.41 0.00   1   2     1
COUNTRY_SATIS          8 250 0.56 0.50      1    0.58 0.00   0   1     1
```

Fig. 12: Descriptive statistics

A correlation plot is also checked for further understanding of the data. Fig. 13 shows the combination of the correlation with linear relation between each variable and correlation values, and distribution plot for each variable in between. The order of the variables in the fig. 13 are: 1. ECON_SIT, 2. CHILDREN_BETTEROFF2, 3. SATISFIED_DEMOCRACY, 4. FUTURE_EDUCATION, 5. MARKET_ECON, 6. FREE_EDUCATION, 7. SEX, 8. COUNTRY_SATIS. All the details discussed above can be visually seen by looking at the distribution plot of each variable. Also, there is no significant, above 0.5, correlation between all of our

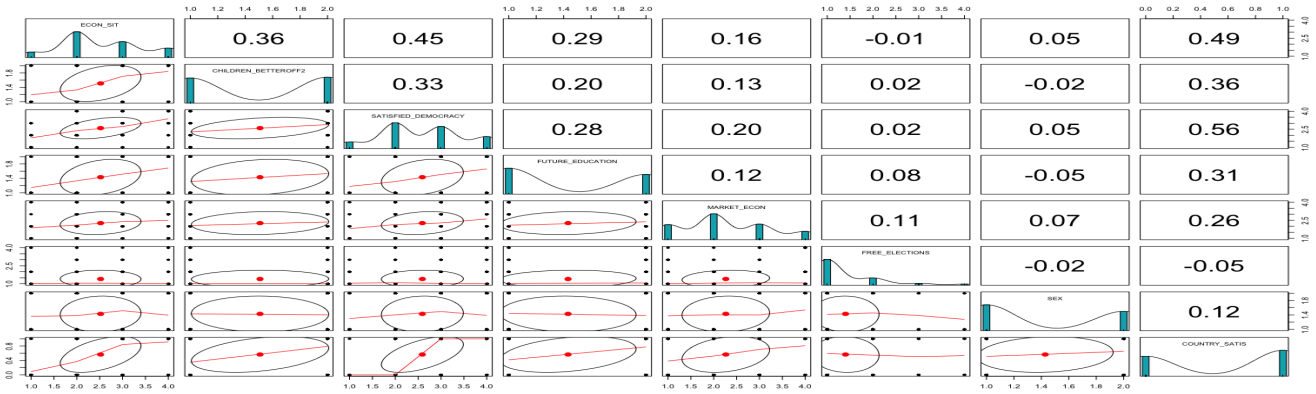


Fig. 13: Correlation plot of the Satisfaction data

independent variables. Thus, it is expected that assumption of no multicollinearity will be satisfied after model building. By looking at linear charts between dependent COUNTRY_SATIS and independent variables, a linear correlation is noted for variables ECON_SIT and SATISFIED_DEMOCRACY. It tells that individuals who are less satisfied with the economic and democracy are dissatisfied with their country in general. It is noted that variable FREE_ELECTION has no linear relationship with COUNTRY_SATIS, thus it might be insignificant while model building.

Thus, by noting all these points from the descriptive statistics performed above, model building is performed on this data.

D. Model Building

In this section, a logistic regression model is built and finalized by removal of insignifican variables based on Wald statistic. Before model building, the data is split using random sample in 80:20 train and test data for further model evaluation purpose. The train set has 200 and test set has 50 out of 250 observations. This train set is used for modeling purpose further.

A model is first built considering all the variables on our dependent variable COUNTRY_SATIS. Fig. 14 shows summary all the models built to finalize the model. Fig. 14a shows the summary of first model. As seen, variable FREE_ELECTION is insignificant for classification of Satisfaction level of an individual. As it was expected from the above descriptive analysis section, this variable is removed. Null and residual deviance for this model is noted as 273.33 and 159.46 respectively. A new model without FREE_ELECTION is fitted to the data. Fig. 14b shows the summary of the new updated model. All the variables are significant except for MARKET_ECON which has p-value closer to 0.1. Thus, this variable is removed. The residual deviance of this model, 162.04. The model is then updated with the remaining variables. Fig. 14c shows the final model built with significant variables. This means all the variables are significant in predicting the probability of an individual being satisfied or not with their country's situation. Thus, this model is finalized for satisfaction classification purpose. The residual deviance is 165.01 which is 108.32 less from the null deviance. Also, by removing the FREE_ELECTION and MARKET_ECON, the residual deviance is increased just by 5.55 which is considerably less. By removing 2 variables to come down to total of 5 independent variables with just 5.55 increase in deviance, principal of parsimony is satisfied. Also, AIC metric is increased by just 1.55 from 175.46 to 177.01, meaning this final model is as good as the initial model in prediction of the dependent variable. The final model includes

```
> satisfaction_model <- glm(COUNTRY_SATIS~., data = ga_sampled_250_train, family = "binomial")
> summary(satisfaction_model)

Call:
glm(formula = COUNTRY_SATIS ~ ., family = "binomial", data = ga_sampled_250_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9302  -0.7123   0.1289   0.5899   2.2305

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -10.6790     1.7423  -6.129 8.83e-10 ***
ECON_SIT             1.0719     0.2958   3.624  0.00029 ***
CHILDREN_BETTEROFF2  0.9844     0.4108   2.397  0.01655 *
SATISFIED_DEMOCRACY  1.5343     0.3133   4.897 9.75e-07 ***
FUTURE_EDUCATION     1.0962     0.4204   2.608  0.00912 **
MARKET_ECON          0.4695     0.2344   2.003  0.04514 *
FREE_ELECTIONS      -0.4275     0.2672  -1.600  0.10965
SEX                  0.9009     0.4172   2.159  0.03082 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 273.33  on 199  degrees of freedom
Residual deviance: 159.46  on 192  degrees of freedom
AIC: 175.46
```

(a) Model with all variables

```
> satisfaction_model_1 <- glm(COUNTRY_SATIS~.-FREE_ELECTIONS, data = ga_sampled_250_train,
family = "binomial")
> summary(satisfaction_model_1)

Call:
glm(formula = COUNTRY_SATIS ~ . - FREE_ELECTIONS, family = "binomial",
    data = ga_sampled_250_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1801  -0.7206   0.1346   0.6033   2.0989

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -10.8805     1.7130  -6.352 2.13e-10 ***
ECON_SIT             1.0903     0.2912   3.744 0.000181 ***
CHILDREN_BETTEROFF2  0.9915     0.4082   2.429 0.015130 *
SATISFIED_DEMOCRACY  1.4964     0.3066   4.880 1.06e-06 ***
FUTURE_EDUCATION     1.0003     0.4094   2.443 0.014550 *
MARKET_ECON          0.3800     0.2229   1.705 0.088242 .
SEX                  0.8698     0.4114   2.114 0.034498 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 273.33  on 199  degrees of freedom
Residual deviance: 162.04  on 193  degrees of freedom
AIC: 176.04

Number of Fisher Scoring iterations: 6
```

(b) Without FREE_ELECTIONS

```
> satisfaction_model_2 <- update(satisfaction_model_1, COUNTRY_SATIS~.-MARKET_ECON)
> summary(satisfaction_model_2)

Call:
glm(formula = COUNTRY_SATIS ~ ECON_SIT + CHILDREN_BETTEROFF2 +
    SATISFIED_DEMOCRACY + FUTURE_EDUCATION + SEX, family = "binomial",
    data = ga_sampled_250_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3209  -0.7172   0.1021   0.6107   2.0987

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -10.2269     1.6181  -6.320 2.61e-10 ***
ECON_SIT             1.1110     0.2903   3.827  0.00013 ***
CHILDREN_BETTEROFF2  1.0203     0.4027   2.534  0.01128 *
SATISFIED_DEMOCRACY  1.5199     0.2993   5.078 3.81e-07 ***
FUTURE_EDUCATION     1.0384     0.4057   2.559  0.01049 *
SEX                  0.8399     0.4044   2.077  0.03783 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 273.33  on 199  degrees of freedom
Residual deviance: 165.01  on 194  degrees of freedom
AIC: 177.01

Number of Fisher Scoring iterations: 6
```

(c) Final Without MARKET_ECON

Fig. 14: Model Summaries

independent variables shown in fig. 14c. For this model to be generalized for the population data, all the assumptions of the Maximum Likelihood method of logistic regression model are checked and discussed in the next section.

E. Diagnostics and Assumptions Check

Maximum likelihood of logistic regression have 5 assumptions that must be followed by the model for it to called generalized and acceptable. Those are defined and discussed for our model as follows:

1) Mutually Exclusive Dependent Variable: As discussed in the Data description, our dependent variable COUNTRY_SATIS has only two values and totally opposite to each other, thus fulfilling this assumption.

2) Reasonably high Sample size: For a maximum likelihood model to converge, it is necessary that the data must have at least minimum of 60 cases or n*20 observations in the data where n is number of variables in the dataset. The data used to build the model has total of 8 variables and 200 cases(training) which fulfills this assumption.

3) Absence of Multicollinearity: If the explanatory variables have high degree of collinearity in them, the method is not accepted to be generalized as it becomes unclear which variable is actually acting to affect the variance of the response variable. Thus, it is assumed that the independent variables are not highly correlated with each other. To check this, VIF measure is used on this model which is called as Variance Inflation Factor which quantifies the multicollinearity in the regression model. Fig. 15 shows the VIF for all the independent variables, and all the variables have vif scores around 1. Thus, our model is following the assumption of absence of multicollinearity. Thus, the expectation of no multicollinearity in the descriptive statistics section discussed is true.

```
> vif(satisfaction_model_2)
          ECON_SIT CHILDREN_BETTEROFF2 SATISFIED_DEMOCRACY    FUTURE_EDUCATION
          1.031628           1.061023           1.037300           1.036159
               SEX
          1.080842
```

Fig. 15: VIF for all independent variables

4) Absence of Outliers: Outliers are influencing observations which have a capacity to throw off a regression model for true regression pattern. Thus, it is necessary that no case in the data is an outlier. This assumption is checked by plotting Cook's distances of all our data points. Fig. 16 shows the visual plot of Cook's distance of our model data. As all the data points have Cook's distance below 0.1, no influential data points is confirmed in our data.

5) Independence of Errors: This assumption focuses on the observations are being related or not to each other. It assumes that the observations in the data are not related to each other and the response variable value for all different cases is independent of each other. Here, we assume that each case has response variable which was noted being independent of the other individual in the dataset.

By above discussed assumption checks on the model, our final model is called to be a generalized acceptable model to perform classification on the population of the given set of data.

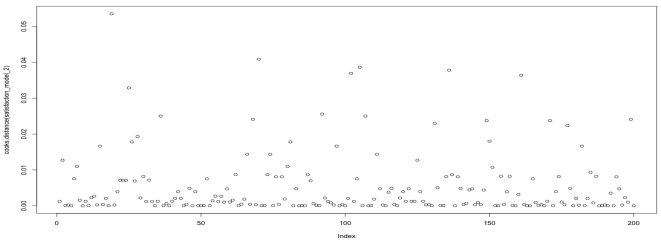


Fig. 16: Cooks distance plot

F. Model Evaluation

To perform model evaluation, the model is used to predict the values on the test data created earlier. This predicted values are then used for evaluation purposes. 0.5 is the cut-off probability used for predicting the response values. To evaluate this final generalized model, 5 metrics are checked and discussed as follows. Five metrics are checked for model evaluation as evaluation of logistic regression model is not simple as multiple linear regression. It is necessary to know how good this model actually is for classification purposes.

1) Hosmer and Lemeshow Goodness-of-fit test: It is first necessary to check if the model is a good or poor fit model to the data provided. If it is poorly fitting model, then it is not a generalized predicting model. This test performs a goodness of fit which performs a step wise, subset of data, comparison between the observed response values and the expected values. For an insignificant p-value confirms that the model is a good fit. Fig. 17 shows the Hosmer and Lemeshow Goodness-of-fit test performed on our model. As the p-value is 0.8384, the model built is good-fit model.

```
> hoslem.test(ga_sampled_250_train$COUNTRY_SATIS, satisfaction_model_2$fitted.values, g=20)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  ga_sampled_250_train$COUNTRY_SATIS, satisfaction_model_2$fitted.values
X-squared = 12.169, df = 18, p-value = 0.8384
```

Fig. 17: Hosmer and Lemeshow Goodness-of-fit test

2) Accuracy measure: A model's performance can be evaluated on the basis of how many cases are correctly or incorrectly predicted. This is measured by accuracy metric from classification cross-table. But, to know how better a model is it is necessary to compare the model accuracy with just a random null model. It shows how better the model is actually when we consider all the variables for prediction than just random prediction. Fig.18 show the null accuracy of the model for test data which is just 54%. This accuracy is now compared with predicted values using our fitted model on the test data. Fig.19 shows the confusion matrix of the predicted values for the test data. The accuracy seen is 82%. Thus, accuracy jumped 28% from 54% null model. This rise in accuracy is significant and a clear indication of our model success in classification of the response variable.

3) Deviance measure: This measure, also called as -2 Log likelihood, is analogous to Residual Standard Error in linear

```
> crosstab   # Null Accuracy = 27/50 = 0.54%
          actual
prediction  0  1
         1 23 27
```

Fig. 18: Null Accuracy of logistic model

```
> confusionMatrix(as.factor(satisfaction_p),
positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 16  2
         1  7 25

               Accuracy : 0.82
                 95% CI : (0.6856, 0.9142)
    No Information Rate : 0.54
    P-Value [Acc > NIR] : 3.251e-05

                  Kappa : 0.6318

 Mcnemar's Test P-Value : 0.1824

            Sensitivity : 0.9259
            Specificity : 0.6957
         Pos Pred Value : 0.7812
         Neg Pred Value : 0.8889
             Prevalence : 0.5400
         Detection Rate : 0.5000
   Detection Prevalence : 0.6400
      Balanced Accuracy : 0.8108

       'Positive' Class : 1
```

Fig. 19: Confusion Matrix of tested model

regression ordinary least square method. Lower the value of residual deviance, better is the accuracy of the model. From the fig. 14c of summary of the model, discussed earlier, the model residual deviance is 165.01 and null deviance is 273.33. It is 108.32 deviance lower than the null model deviance. Given the set of variables in this data, the minimum residual deviance that could be achieved was 159.46 as discussed earlier and shown in the fig. 14. The final model residual deviance is just 5.55 more than that with a better fit.

4) Pseudo R^2 metric: R^2 metric is used to check the amount of variation of the response variable influenced by the inclusion of explanatory variables in the linear regression. In logistic regression, an analogous metric to R^2 which is pseudo R^2 is calculated to get the similar information of the relationship explanation. Fig.20 shows the 3 R^2 values for our data. As seen, the McFadden R^2 is 0.39, the Cox and Snell R^2 is 0.41, and Nagelkerke R^2 value is 0.56. For the given set of variables, these R^2 values are an indication of a good relationship between the explanatory variables and the response variable.

```
> nagelkerke(satisfaction_model_2)
$Models

Model: "glm, COUNTRY_SATIS ~ ECON_SIT + CHILDREN_BETTEROFF2 + SATISFIED_DEMOCRACY + FUTURE_
EDUCATION + SEX, binomial, ga_sampled_250_train"
Null:  "glm, COUNTRY_SATIS ~ 1, binomial, ga_sampled_250_train"

$Pseudo.R.squared.for.model.vs.null
                            Pseudo.R.squared
McFadden                           0.396271
Cox and Snell (ML)                 0.418157
Nagelkerke (Cragg and Uhler)       0.561259
```

Fig. 20: R^2 Measures

5) Sensitivity and Specificity: Sensitivity measure calculates correctly classified true positives from all the observed true positive values. Specificity measure calculates correctly classified true negatives from all the observed true negative values. From the fig. 19, the sensitivity for the test prediction is 0.92% and specificity is 0.69%. It is interpreted as for all the individuals who are truly dissatisfied with the current situation of their country, 92% of them are correctly classified as dissatisfied. Whereas, for those who are truly satisfied, only 69% of them are correctly classified as satisfied and 31% of them are classified as dissatisfied by our model on the test data.

From above evaluation of our final generalized logistic regression model, it is seen that our model is a good-fitted model with 28% increased accuracy compared to a random null model with minimum residual deviance and a reasonable pseudo R^2 values showing good relationship between response variable and explanatory variables considered in the final model. Thus, the final model can be considered as a good model for use for the classification of the individual satisfaction level for the population of the given data.

G. Conclusion and Summary of the Model

The final model built have independent variables ECON_SIT, CHILDREN_BETTEROFF2, SATISFIED_DEMOCRACY, FUTURE_EDUCATION, SEX. By performing all the required diagnostic checks, all the assumptions of Maximum likelihood method are checked and the final model is called to be acceptable generalized model for prediction on population data. This generalized model is then evaluated on the test data. By performing Hosmer & Lemeshow Goodness-of-fit test, the model shows a good-fit nature. It is concluded that the model performance compared to random model is considerably good with 82% accuracy and 165.01 residual deviance. Fig. 21 shows the coefficients of the variables of the model. By inclusion of these variables the equation of the logistic regression model is given as

$$E(y) = \frac{e^{-10.226+1.11x_1+1.02x_2+1.51x_3+1.03x_4+0.83x_5}}{1+e^{-10.226+1.11x_1+1.02x_2+1.51x_3+1.03x_4+0.83x_5}} \tag{1}$$

As seen, it is difficult to interpret the equation to understand the influence of each variable on the response variable. Thus, the model is interpreted using odds ratio of these variables which are exponents of their coefficients. Odds ratio of the variables can be compared directly to know which variables have more influence on the response variable and otherwise. Odds ratio of a variable is the change in odds where the response variable is shifted from 0 to 1 or 1 to 0 by 1 unit increase in explanatory variable. Fig. 22 shows the Odds Ratio of all the independent variables of our logit model. Interpretation of odds ratio of each variable is discussed as follows:

```
> coef(satisfaction_model_2)
       (Intercept)            ECON_SIT CHILDREN_BETTEROFF2
       -10.2268825           1.1110263           1.0203299
SATISFIED_DEMOCRACY    FUTURE_EDUCATION                 SEX
         1.5199270           1.0383838           0.8398791
```

Fig. 21: Coefficients of the final logistic model

```
> # Odds Ratios
> exp(coef(satisfaction_model_2))
       (Intercept)            ECON_SIT CHILDREN_BETTEROFF2
      0.0000361844        3.0374741855        2.7741097458
SATISFIED_DEMOCRACY    FUTURE_EDUCATION                 SEX
      4.5718913078        2.8246481051        2.3160870073
```

Fig. 22: Odds ratios of variables

1) ECON_SIT: The odds of an individual to be dissatisfied with their country is 3.03 times higher who thinks that the economic situation of the country is "Very bad" than those

who thinks a one unit less of "Very bad". If the answer jumps from 1, "Very good", to 3, "Somewhat bad", the odds that the person will be dissatisfied with his or her country will be 3.03^2

2) CHILDREN_BETTEROFF2: For an individual having opinion "Worse off" for children having better future, the odds of that person being dissatisfied is 2.77 times than those who have opinion as "Better off".

3) SATISFIED_DEMOCRACY: The odds that the person will be dissatisfied with the country is 4.57 times for those who are "Somewhat satisfied" with the democracy than those who are "Very satisfied".

4) FUTURE_EDUCATION: The odds of a person being dissatisfied is 2.82 times higher for those who are "Optimistic" about the future education of the country than those who are "Pessimistic"

5) SEX: Odds of a person being dissatisfied increases by the factor of 2.31 if it's a "Female" than those for "Male".

Thus, after significant model building, diagnostic checks, model evaluation and interpretation, the target of building a generalized logistic regression model is fulfilled.

IV. Principal Component Analysis

Principal Component Analysis(PCA) is a technique which performs dimension reduction on a data having considerably more variables which are correlated. It reduces the dimension to smaller number of components which are linear combinations of all the variables. The process takes place such that each component explains a certain amount of variability of the entire dataset and no component explains the same variability. Thus, all the components formed are uncorrelated with each other. To explain PCA and steps taken to perform it, an example of superconductor data is considered. PCA is performed using R.

A. Dataset

The Superconductor dataset consists of total 42 variables and 21263 superconductors(rows). Final variable in the dataset is meant for prediction using linear regression using the rest of the variables, thus we discard it in this analysis. Therefore, with 9 main chemical properties of the superconductors like Number of Elements, Atomic Mass, First Ionization Energy(fie), Atomic Radius, Density, Electron Affinity, Fusion Heat, Thermal Conductivity, Valence, and addition of statistical measures of these properties such as mean, geometric mean, entropy, range and standard deviation collectively accounts for a total of 41 variables.

B. Performing PCA on Superconductor data

A step-wise explanation of PCA on the dataset is given as follows:

1) Check for Suitability of the data:

1) A simple Correlation Matrix overview intuition:
   - The variables must be considerably correlated($>$0.3) with each other to be combined as linear combinations to form factors.
   - From the Correlation Matrix in the fig. 23, we can see that many variables are correlated to each other.

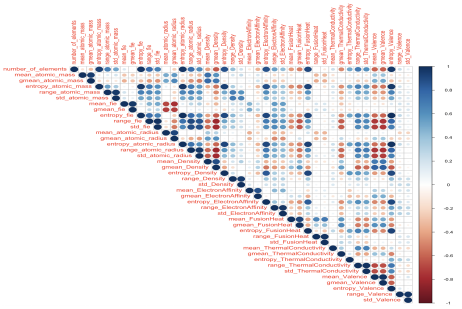


Fig. 23: Correlation Visual plot between variables

2) Bartletts Test of Sphericity:
   - This hypothesis test assumes that correlation between variables are zero. It is useful when we have significantly high number of variables and is difficult to interpret correlation matrix to know correlations between the variables.
   - For the superconductor dataset, we can see from the fig. 24 that the p-value is 2.2e-16, i.e. $<$0.05. Thus, correlation. between variables is confirmed.

```
> bartlett.test(superconductor_data_new)

        Bartlett test of homogeneity of variances

data:  superconductor_data_new
Bartlett's K-squared = 7362891, df = 40, p-value < 2.2e-16
```

Fig. 24: Bartletts Test of Sphericity

3) Kaiser-Meyer-Olkin(KMO) Measure of Sampling Adequacy(MSA score):
   - The score ranging from 0-1, this measure indicates if a given correlation matrix is appropriate for factor analysis.
   - A value $>$0.6 can be an indicator of performing factor analysis.
   - From fig. 25, the KMO Overall MSA score is 0.83, which is in the meritorious range according to KMO [1]. Therefore, we now perform PCA on this data.

```
> KMO(superconductor_data_new)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = superconductor_data_new)
Overall MSA =  0.83
```

Fig. 25: KMO Measure of Sampling Adequacy

2) Factors/Components Extraction:

- Performing PCA on a dataset actually gives same number of components as the number variables in the dataset. But, the underlying variables are linearly combined to get these components. A few components explains majority of the variance in the data. To extract these few components explaining large portion of the data, we use Eigen Value rule.
- Eigen Value(EV) rule, also called as Kaiser's Criterion, says that components having EV $>$1 are the important components and the combination of these components

together explains majority of the variability in the dataset. Rest components are not as useful as these selected components.

- The importance of each component can be visualize using Scree Plot of EV of principal components. For our example superconductor dataset, Scree plot is show in fig. 26 after performing PCA.
- The red dashed line is the EV threshold of 1., Thus, 8 Principal Components have EV above the red dashed line are important.
- Thus, these 8 are selected as our final PCs for our dataset.

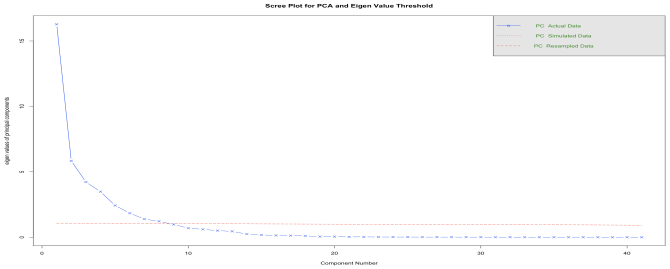


Fig. 26: Scree plot of Eigen Values of each component

3) Principal Component Loadings:

- Understanding which variables to what extent are correlated with each of the factors, we check loadings of each variable on each PCs. In fig. 27, for simplicity of understanding, a visualisation of loadings is plot.
- It is seen that most of the variables are highly correlated with PC1 and then with PC2, PC3, and so on, diminishing in nature.
- It is difficult to interpret which components are actually a base factor on which we can interpret the linear combinations of the variables.
- These 8 PCs cumulatively explains 89% of the total variance in the dataset. Thus, 89% of the variability is retained by reducing dimension significantly from 41 to 8 variables.

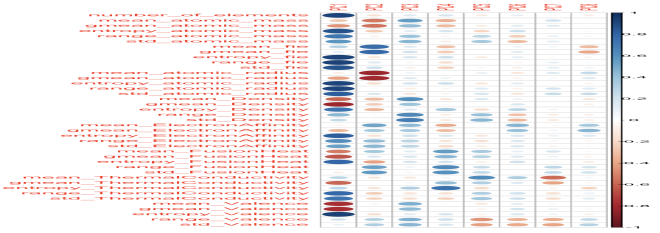


Fig. 27: Variable Loadings on PCs Correlation plot

4) Rotation of the Principal Components:

- To understand which variables are highly correlated and which are not with individual PCs, we use Factor Rotation. Varimax Orthogonal axes rotation is used for this purpose.
- In fig. 28, we can now see a good mix of variables for each of the rotated components which are easier to interpret than non-rotated components.
- Naming these components can be subjective, but as a common interpretation, we can name each components for the superconductor as follows: i. RC1:Entropy combination of all main features. ii. RC6:Mix of density, valence, fie and thermal conducivity. iii. RC2:FIE. iv. RC3:Density Mass. v. RC4:Fusion heat. vi. RC5:Valence. vii. RC8:Election Affinity. viii. RC7:Thermal Conductivity.

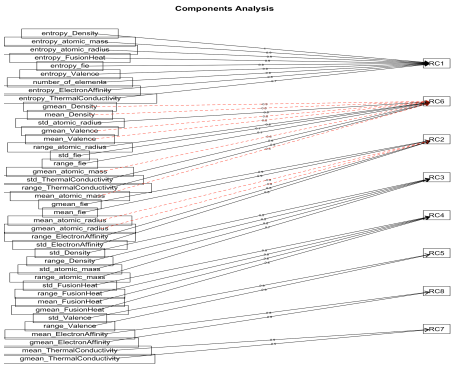


Fig. 28: Variable loadings after Rotation

5) Communalities:

- Communalities table shows, out of the total variance(1), contribution of each variable's explained variance in the final PCs.
- From fig. 29, we can see that, 21 variables have more than 90%, 18 variables have more than 80% and 2 variables have 73% and 65% of variance explained by selecting 8 PCs.

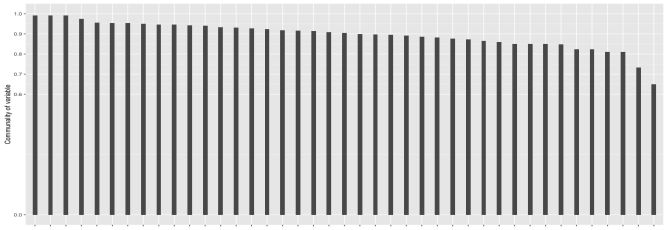


Fig. 29: Communalities of each variable in the Components

Thus, PCA is performed on the superconductor dataset, and all the necessary steps were taken for the evaluation and interpretation of the best PCs selected covering most of the variability of the dataset.

V. Conclusion

A complete statistical analysis is thus performed for time series data, binary classification data using logistic regression and a highly complex data using principal component analysis.

References

[1] Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. Educational and psychological measurement, 34(1), 111-117.

[2] Hyndman RJ. Forecasting: Principles and Practice. 2nd. Otexts; 2018. Accessed January 1, 2021.

[3] Lind DA, Marchal WG, Wathen SA. Statistical Techniques in Business & Economics. 18th. McGraw-Hill Education; 2020. Accessed January 1, 2021.