Binary Logistic Regression and Time Series Analysis

Srijon Datta
MSc. in Data Analytics, Sept'22
x21225265@student.ncirl.ie

***Abstract*—This report presents two different analytical studies on two individual datasets. Firstly, we fit a binary logistic regression model and evaluate the model performance and accuracy based on the marketing campaign data of a certain bank. Here our main objective is to do a thorough discussion on the final model building steps and the reasons for rejecting intermediate models. We also verify if the model assumptions have been satisfied for the final model. Secondly, we perform a time series analysis on the flight departure data starting from the year 2010 to 2022 in Ireland. Here, we assess the nature of the series with appropriate visualizations. Furthermore, we discuss the various diagnostic tests to different time series components and then we try to figure out the best fit model to forecast, discussing the proper justification for rejecting the intermediate models. In this study we will confine ourselves in ARIMA, SARIMA, Exponential Smoothing and basic time series models. We will be using R language for the entire work.**

I. Binary Logistic Regression Analysis

Like all other regression analysis, binary logistic regression analysis is also a predictive analysis. Unlike the linear regression analysis, here the dependent variable is of *dichotomous* (i.e binary) type but the independent variables can be nominal, ordinal, interval or ratio type. The main objective of binary logistic regression is to explain the relationship between one dependent binary variable and the other independent variables and on that basis make a predictive model to predict future responses on the similar type of data. In this study, a binary logistic regression model has been tried to build up based on the following data-set and proper discussions of model building steps and different accuracy measures has been stated.

A. Data Description

This analysis is done on the *bank data* which contains details of a marketing campaign that aims to convince the customer to buy a bank product. The data consists of 17 variables and 45,211 samples. All the variable descriptions are given in Table I.

B. Objective

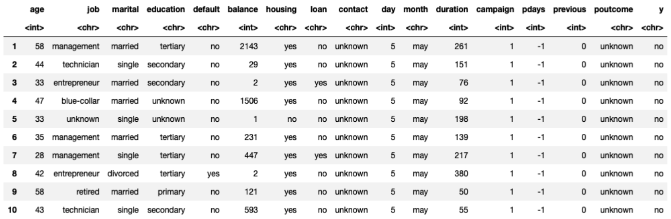
In this work, the prime objective is to estimate a binary logistic regression model to facilitate understanding of the relationships between different Marketing campaign characteristics and classification as 'yes' or 'no' based on the above mention data. We also want to discuss the model-building steps that has been undertaken to arrive at the final logistic regression model and the reasons for rejecting intermediate models clearly. Finally we want to check the final model performance accuracy and the visual representation of the model performance.

TABLE I: Variable Description Table

Variable Names	Descriptions
age	age of the customers
job	what do the customers do
marital status	marital status of the customers
education	education levels of the customers
credit	do the customers have ongoing credit
housing	customers have mortgage or not
loan	customers has personal loans or not
contact	communication type with customers
day	last contact day of the week
month	last contact month
duration	last contact duration, in seconds (numeric)
campaign	number of contacts performed during this campaign and for this customer
pdays	number of days that passed by after the client was last contacted
previous	number of contacts performed before this campaign and for this client
poutcome	outcome of the previous marketing campaign
y	has the client subscribed for the bank product

C. Descriptive Statistics & Data Visualization

Figure 1 refers the first 10 samples of the data-set. Here we can observe that there is a variable name *default* but in variable description we have no such variable. So by judging the nature of this variable we assume the variable *credit* in place of that and while data pre-processing we should renamed the variable to credit.



	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
	<int>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<int>	<chr>	<int>	<int>	<int>	<int>	<chr>	<chr>
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no

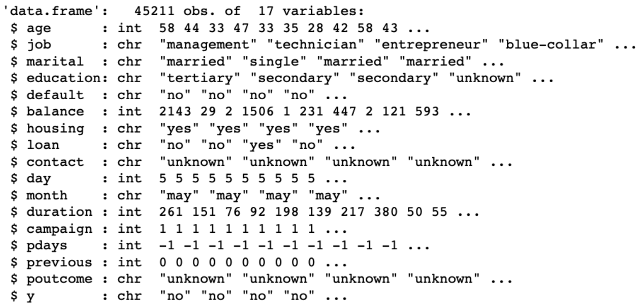
Fig. 1: First 10 rows of the data

Figure 2 refers the structures of the data. Here we can clearly observe that all the *character variables* (except the month variable) contains different factor levels. So we should consider the variable types for all the character variables to factors.

Figure 3 shows the descriptive statistics of all the numeric variables present in the data. The minimum, third quartile and the maximum values for each of the numeric data suggest that there is a high chance of outliers being present in the data.

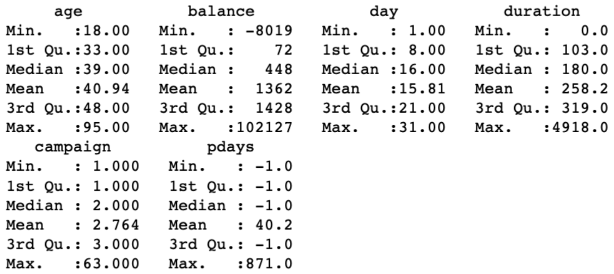
D. Data Pre-processing before Model-1

We have already observed some requirements of data pre-processing in the above section. *Firstly*, we have renamed the variable default to credit as we have noticed that there is no variable named default in the variable descriptions. *Secondly*, we have changed all the character type variables to factor type variables (except month) as they contains different factor levels of the corresponding variables. *Finally*, we have checked whether there is an missing values or NA values in the whole data-set or not and found that the data is clean.



```
'data.frame':   45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital  : chr  "married" "single" "married" "married" ...
 $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
 $ default  : chr  "no" "no" "no" "no" ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : chr  "yes" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "no" "yes" "no" ...
 $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : chr  "may" "may" "may" "may" ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ y        : chr  "no" "no" "no" "no" ...
```

Fig. 2: Structure of the data



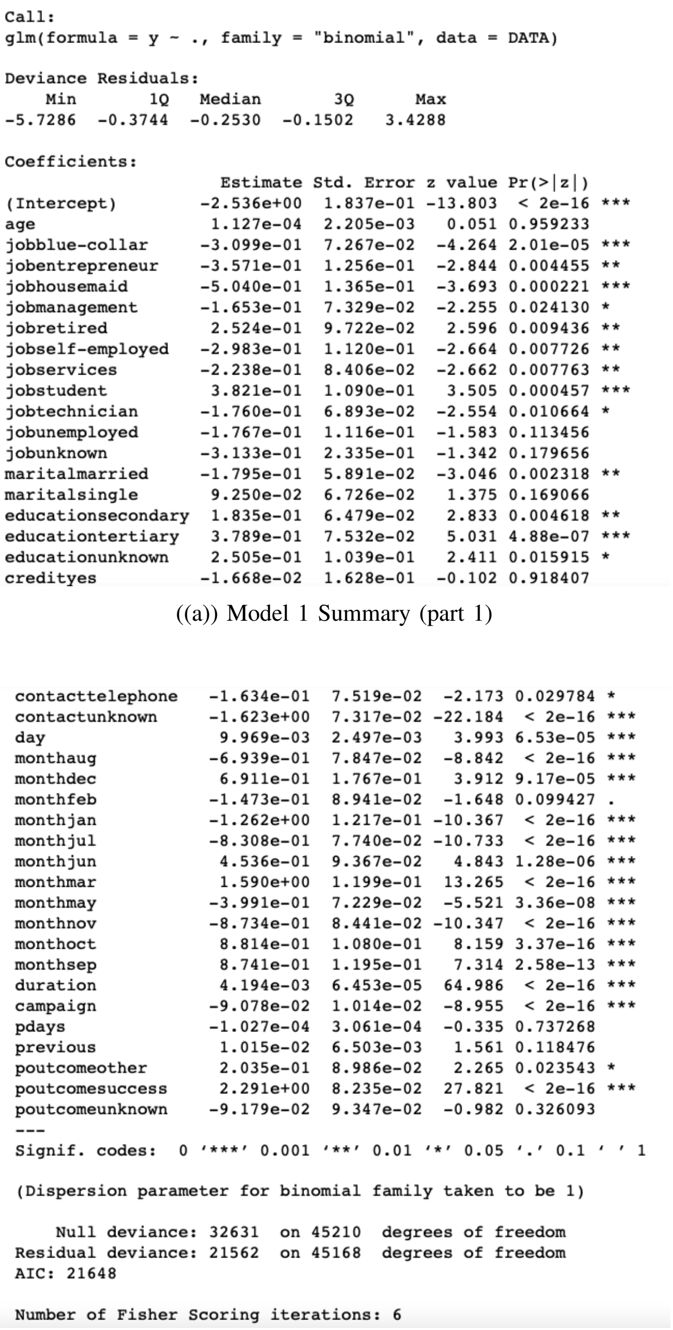
```
      age           balance            day           duration     
 Min.   :18.00   Min.   : -8019   Min.   : 1.00   Min.   :   0.0  
 1st Qu.:33.00   1st Qu.:    72   1st Qu.: 8.00   1st Qu.: 103.0  
 Median :39.00   Median :   448   Median :16.00   Median : 180.0  
 Mean   :40.94   Mean   :  1362   Mean   :15.81   Mean   : 258.2  
 3rd Qu.:48.00   3rd Qu.:  1428   3rd Qu.:21.00   3rd Qu.: 319.0  
 Max.   :95.00   Max.   :102127   Max.   :31.00   Max.   :4918.0  
    campaign          pdays      
 Min.   : 1.000   Min.   : -1.0  
 1st Qu.: 1.000   1st Qu.: -1.0  
 Median : 2.000   Median : -1.0  
 Mean   : 2.764   Mean   : 40.2  
 3rd Qu.: 3.000   3rd Qu.: -1.0  
 Max.   :63.000   Max.   :871.0  
```

Fig. 3: Descriptive Statistics of numeric variables

E. Model-1

This is a naive binary logistic regression model since all the variables have been incorporated in the model and no outliers have been removed from the data-set. We have performed this model for the initial inspection purpose and tried to check how does binary logistic regression model performs in this data.

In Figure 4(a), the deviance residuals have minimum of -5.72, median of -0.25 and the maximum of 3.42 which suggests that instead of symmetric about zero, the deviance residuals are slightly right skewed. Again it can be clearly seen that some of the p-values of some variables' coefficients are significantly larger than the level of significance (in our case 0.05) suggesting the insignificant affect of those variables on the target value in the model. So we need to filter those insignificant variables before further modelling. Figure 4(b) refers that the residual deviance is 21,562 which is considerable large and the Akaike Information Criterion (AIC) value is 21,648 for this model. AIC value is used to compare multiple models. Comparatively smaller AIC value suggests a good fitted model.



```
Call:
glm(formula = y ~ ., family = "binomial", data = DATA)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.7286  -0.3744  -0.2530  -0.1502   3.4288  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -2.536e+00  1.837e-01 -13.803  < 2e-16 ***
age                 1.127e-04  2.205e-03   0.051 0.959233    
jobblue-collar     -3.099e-01  7.267e-02  -4.264 2.01e-05 ***
jobentrepreneur    -3.571e-01  1.256e-01  -2.844 0.004455 ** 
jobhousemaid       -5.040e-01  1.365e-01  -3.693 0.000221 ***
jobmanagement      -1.653e-01  7.329e-02  -2.255 0.024130 *  
jobretired          2.524e-01  9.722e-02   2.596 0.009436 ** 
jobself-employed   -2.983e-01  1.120e-01  -2.664 0.007726 ** 
jobservices        -2.238e-01  8.406e-02  -2.662 0.007763 ** 
jobstudent          3.821e-01  1.090e-01   3.505 0.000457 ***
jobtechnician      -1.760e-01  6.893e-02  -2.554 0.010664 *  
jobunemployed      -1.767e-01  1.116e-01  -1.583 0.113456    
jobunknown         -3.133e-01  2.335e-01  -1.342 0.179656    
maritalmarried     -1.795e-01  5.891e-02  -3.046 0.002318 ** 
maritalsingle       9.250e-02  6.726e-02   1.375 0.169066    
educationsecondary  1.835e-01  6.479e-02   2.833 0.004618 ** 
educationtertiary   3.789e-01  7.532e-02   5.031 4.88e-07 ***
educationunknown    2.505e-01  1.039e-01   2.411 0.015915 *  
credityes          -1.668e-02  1.628e-01  -0.102 0.918407    
```

((a)) Model 1 Summary (part 1)

```
contacttelephone   -1.634e-01  7.519e-02  -2.173 0.029784 *  
contactunknown     -1.623e+00  7.317e-02 -22.184  < 2e-16 ***
day                 9.969e-03  2.497e-03   3.993 6.53e-05 ***
monthaug           -6.939e-01  7.847e-02  -8.842  < 2e-16 ***
monthdec            6.911e-01  1.767e-01   3.912 9.17e-05 ***
monthfeb           -1.473e-01  8.941e-02  -1.648 0.099427 .  
monthjan           -1.262e+00  1.217e-01 -10.367  < 2e-16 ***
monthjul           -8.308e-01  7.740e-02 -10.733  < 2e-16 ***
monthjun            4.536e-01  9.367e-02   4.843 1.28e-06 ***
monthmar            1.590e+00  1.199e-01  13.265  < 2e-16 ***
monthmay           -3.991e-01  7.229e-02  -5.521 3.36e-08 ***
monthnov           -8.734e-01  8.441e-02 -10.347  < 2e-16 ***
monthoct            8.814e-01  1.080e-01   8.159 3.37e-16 ***
monthsep            8.741e-01  1.195e-01   7.314 2.58e-13 ***
duration            4.194e-03  6.453e-05  64.986  < 2e-16 ***
campaign           -9.078e-02  1.014e-02  -8.955  < 2e-16 ***
pdays              -1.027e-04  3.061e-04  -0.335 0.737268    
previous            1.015e-02  6.503e-03   1.561 0.118476    
poutcomeother       2.035e-01  8.986e-02   2.265 0.023543 *  
poutcomesuccess     2.291e+00  8.235e-02  27.821  < 2e-16 ***
poutcomeunknown    -9.179e-02  9.347e-02  -0.982 0.326093    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 21562  on 45168  degrees of freedom
AIC: 21648

Number of Fisher Scoring iterations: 6
```

((b)) Model 1 Summary (part 2)

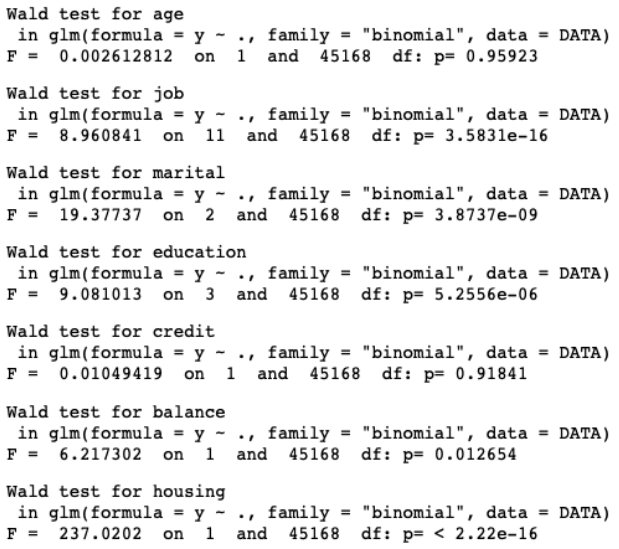
Fig. 4: Model 1 Summary

F. Further Data Pre-processing before Model 2

After model 1, we have performed the following tests to identify the variables of interests for our binary logistic model -

1) Wald test [1]: A wald test is used to evaluate the statistical significance of each coefficient in the model and is calculated by taking the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The idea is to test the hypothesis that the coefficient of an independent variable in the model is significantly different from zero. If the test fails to reject the null hypothesis, this suggests that removing the variable from the model will not substantially harm the fit of that model.

So we have conducted wald test for all the independent variables (Figure5) and found that some of the variables like *age, credit, pdays, previous* have insignificant affect in our model since the large p-value of the test fails to reject the null hypothesis at 5% level of significance. So we have removed those insignificant variables from the model data before jump into the next logistic model.



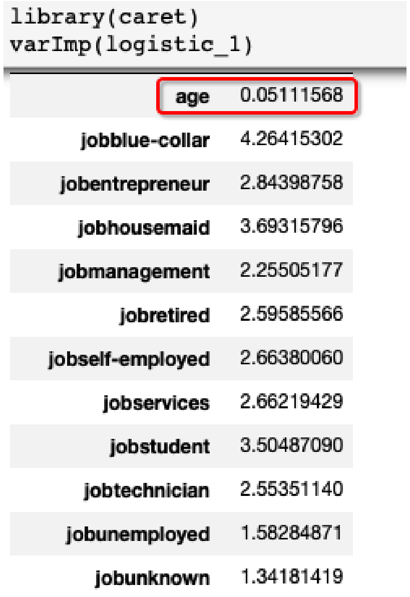
```
Wald test for age
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  0.002612812  on  1  and  45168  df: p= 0.95923

Wald test for job
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  8.960841  on  11  and  45168  df: p= 3.5831e-16

Wald test for marital
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  19.37737  on  2  and  45168  df: p= 3.8737e-09

Wald test for education
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  9.081013  on  3  and  45168  df: p= 5.2556e-06

Wald test for credit
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  0.01049419  on  1  and  45168  df: p= 0.91841

Wald test for balance
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  6.217302  on  1  and  45168  df: p= 0.012654

Wald test for housing
 in glm(formula = y ~ ., family = "binomial", data = DATA)
F =  237.0202  on  1  and  45168  df: p= < 2.22e-16
```

Fig. 5: Wald Test

2) Variable Importance: To assess the relative importance of individual predictors in the model, we can also look at the absolute value of the t-statistic for each model parameter. This technique is utilized by the 'varImp' function in the caret package in R for general and generalized linear models. A significant smaller value for t-statistic also suggests us the insignificant affect of that particular variable on the target variable in our model. Figure 6 depicts that the variable *age* has comparatively smaller t-value than others and hence the removal of above mentioned insignificant variables is justified.

G. Model 2

Here we have tried to fit a binary logistic regression model with the variables of interest only and we have not found much improvement result. Though the residual deviance and the AIC values have reduced to 21,508 and 21,586 respectively but the reductions are not significant. Through this model we have successfully predicted 38,933 and 1,840 observations as true negatives and true positives respectively among 45,208 observations, achieving approximately 90% model accuracy.



```
library(caret)
varImp(logistic_1)
```

age	0.05111568
jobblue-collar	4.26415302
jobentrepreneur	2.84398758
jobhousemaid	3.69315796
jobmanagement	2.25505177
jobretired	2.59585566
jobself-employed	2.66380060
jobservices	2.66219429
jobstudent	3.50487090
jobtechnician	2.55351140
jobunemployed	1.58284871
jobunknown	1.34181419

Fig. 6: Variable Importance

Judging by the residual vs fitted plot and scale location curve, in Figure 7, we can understand that there are a lot of influential outliers in the data-set. Therefore we need to handle those outliers before further modelling to improve our model outcomes.

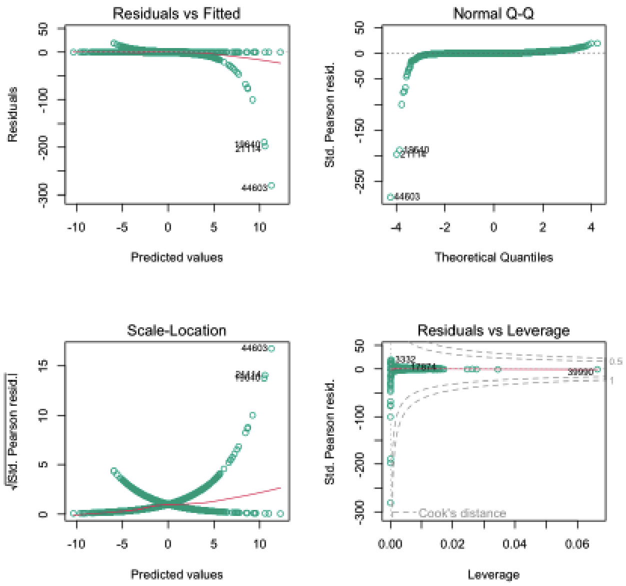


Fig. 7: Visualization of Model 2

H. Identifying and Removing Outliers

Figure 8 suggests that there are a large number of outliers in some of the numeric independent variables.

We have used *Inter-Quarter-Range* (IQR) technique to identify the outliers in the data-set. In this technique we

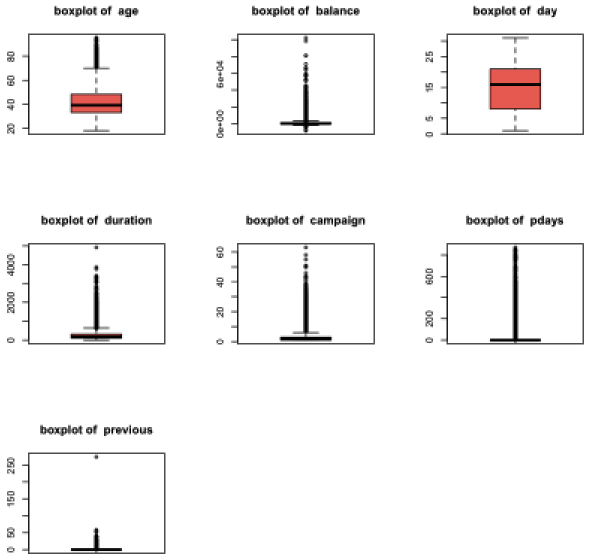
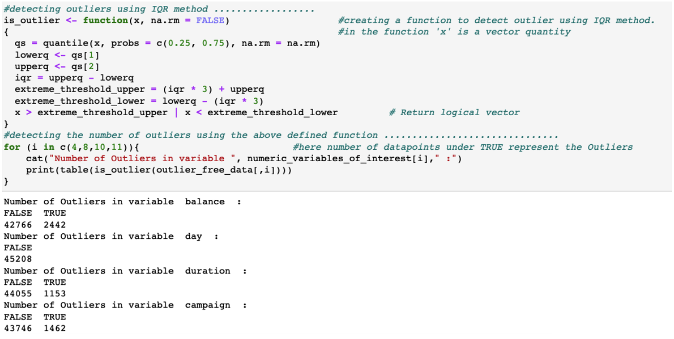


Fig. 8: Box-plot of Numeric Variables

usually calculate the inter-quarter range and find out those observations which are smaller than the difference of first quarter and the 3 times of the IQR and larger than the difference of the third quarter and the 3 times of the IQR for a particular variable. We have defined a function based on this technique (see Figure 9) to identify the influential outliers in each numeric variables (since categorical variables can not have outliers) and removed those from the model.



```
#detecting outliers using IQR method ..................
is_outlier <- function(x, na.rm = FALSE)                    #creating a function to detect outlier using IQR method.
{                                                           #in the function 'x' is a vector quantity
  qs = quantile(x, probs = c(0.25, 0.75), na.rm = na.rm)
  lowerq <- qs[1]
  upperq <- qs[2]
  iqr = upperq - lowerq
  extreme_threshold_upper = (iqr * 3) + upperq
  extreme_threshold_lower = lowerq - (iqr * 3)
  x > extreme_threshold_upper | x < extreme_threshold_lower          # Return logical vector
}
#detecting the number of outliers using the above defined function ..............................
for (i in c(4,8,10,11)){                                 #here number of datapoints under TRUE represent the Outliers
    cat("Number of Outliers in variable ", numeric_variables_of_interest[i]," :")
    print(table(is_outlier(outlier_free_data[,i])))
}
```

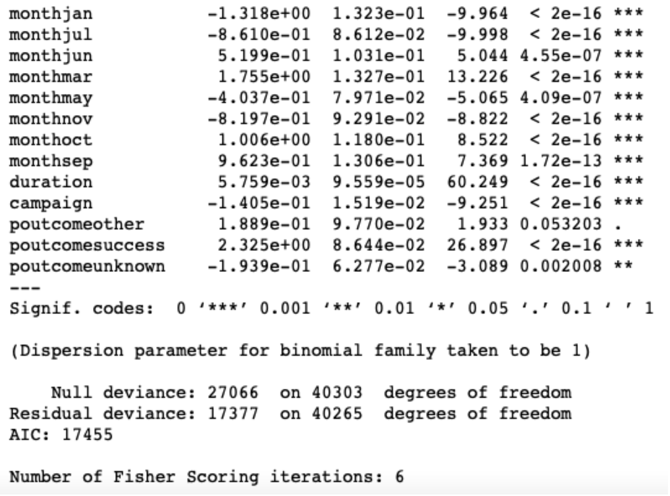
```
Number of Outliers in variable  balance  :
FALSE  TRUE
42766  2442
Number of Outliers in variable  day  :
FALSE
45208
Number of Outliers in variable  duration  :
FALSE  TRUE
44055  1153
Number of Outliers in variable  campaign  :
FALSE  TRUE
43746  1462
```

Fig. 9: Identifying Outliers

Before removing the outliers we had 45,208 sample data but after removing the outliers we have in total 40,304 samples left. So 4,904 outliers have been removed from out data-set.

I. Model 3

Here we have constructed a binary logistic regression model using the outlier treated data. From the model summary (see Figure 10) we can see that the residual deviance has reduced to 17,377 from 21,508 and the AIC value has also significantly reduced to 17,455 from 21,568 as compared to the model 2.



```
monthjan         -1.318e+00  1.323e-01  -9.964  < 2e-16 ***
monthjul         -8.610e-01  8.612e-02  -9.998  < 2e-16 ***
monthjun          5.199e-01  1.031e-01   5.044 4.55e-07 ***
monthmar          1.755e+00  1.327e-01  13.226  < 2e-16 ***
monthmay         -4.037e-01  7.971e-02  -5.065 4.09e-07 ***
monthnov         -8.197e-01  9.291e-02  -8.822  < 2e-16 ***
monthoct          1.006e+00  1.180e-01   8.522  < 2e-16 ***
monthsep          9.623e-01  1.306e-01   7.369 1.72e-13 ***
duration          5.759e-03  9.559e-05  60.249  < 2e-16 ***
campaign         -1.405e-01  1.519e-02  -9.251  < 2e-16 ***
poutcomeother     1.889e-01  9.770e-02   1.933 0.053203 .
poutcomesuccess   2.325e+00  8.644e-02  26.897  < 2e-16 ***
poutcomeunknown  -1.939e-01  6.277e-02  -3.089 0.002008 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27066  on 40303  degrees of freedom
Residual deviance: 17377  on 40265  degrees of freedom
AIC: 17455

Number of Fisher Scoring iterations: 6
```

Fig. 10: Model 3 summary

The residual vs predicted chart for this model has been normalised and the residual vs leverage plot is now symmetric about zero (in Figure 11(a)). Through this model we have successfully predicted 35,248 and 1,466 observations as true negatives and true positives respectively among 40,304 observations (see Figure 11(b)), achieving approximately 91% model accuracy. To verify the model performance, we have calculated the *McFadden's Pseudo* R^2 value (see Figure 11(c)). This gives an equivalent measure like traditional R^2 measure from OLS. Values of 0.2 to 0.4 for McFadden's Pseudo R^2 represent excellent fit for binary logistic regression model [2] and in our case the value is 0.35.

J. Visualization

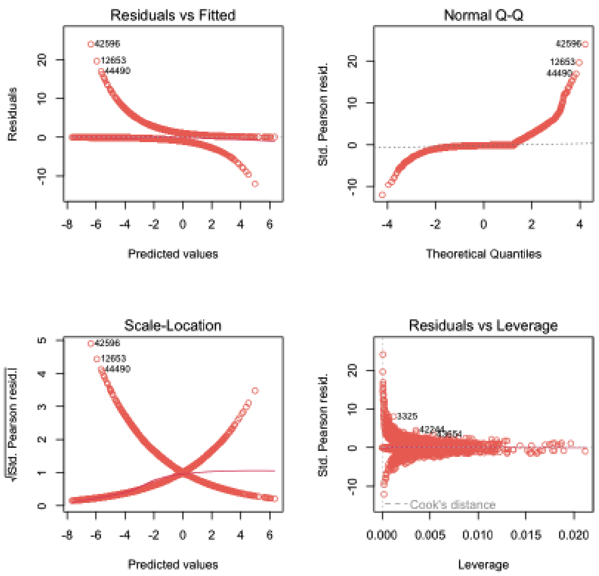
In Figure 12 we have plot the predicted probabilities (in ascending order) of the customers taking the bank's product subscription in y-axis and the corresponding rank of the customer in x-axis.

K. Checking Assumptions for Model 3

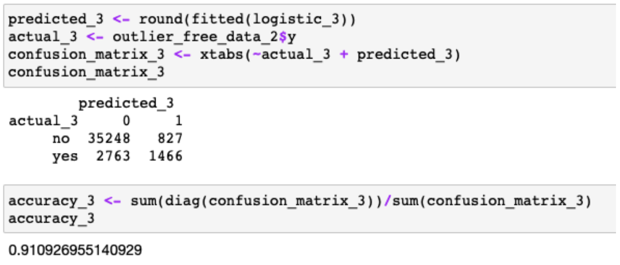
1) Multicollinearity: If two or more predictor variables are found to be highly linearly related to each other in a multiple logistic regression model, then Multicollinearity is said to be present in that model. Presence of multicollinearity could overfit the model performance. To check for multi-collinearity in the independent variables, the Variance Inflation Factor (VIF) technique is used. The variables with VIF score more than 10 means that they are very strongly correlated. Therefore, they are discarded and excluded in the logistic regression model.

In the Figure 13, it is clear that the VIF score for all the explanatory variables are significantly less and hence we can conclude that there is no multicollinearity present in our binary logistic regression model.

2) Linearity: The continuous variables should have linearity against the log(odds) of the dependent variable. To verify this assumption, we have created the predicting probabilities and the logit variable. Then we have tried to visualize the relationship between them by scatter plot. Figure 14 shows that the continuous independent variables are sort of linearly related to the log(odds) of the dependent variable.



((a)) Model 3 plot



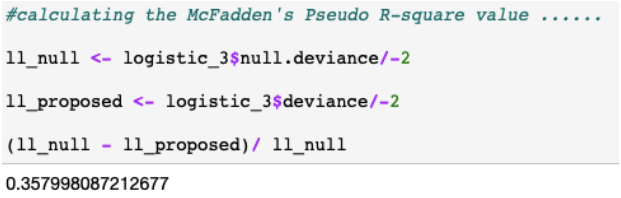
```
predicted_3 <- round(fitted(logistic_3))
actual_3 <- outlier_free_data_2$y
confusion_matrix_3 <- xtabs(~actual_3 + predicted_3)
confusion_matrix_3
```

```
         predicted_3
actual_3      0     1
     no   35248   827
     yes   2763  1466
```

```
accuracy_3 <- sum(diag(confusion_matrix_3))/sum(confusion_matrix_3)
accuracy_3
```

```
0.910926955140929
```

((b)) Model 3 Confusion Matrix and Accuracy



```
#calculating the McFadden's Pseudo R-square value ......

ll_null <- logistic_3$null.deviance/-2

ll_proposed <- logistic_3$deviance/-2

(ll_null - ll_proposed)/ ll_null
```

```
0.357998087212677
```

((c)) McFadden's Pseudo R^2

Fig. 11: Model 3 Performance

3) Outlier Free Data: The data should not possess any outliers in it. Here we measured cook's distance to check the presence of possible outliers in the model data-set. Though we have removed outliers using IQR method but the Figure 15 suggests that few of them are present in the data-set. But we have not removed those sample data due to the possible loss of information for modelling.

L. Conclusion

By all means, model 3 has performed very well in the given data and it satisfies all the assumptions of a good binary logistic regression model. So in our comparative study, model 3 should be considered as the best fitted model for the given data-set.

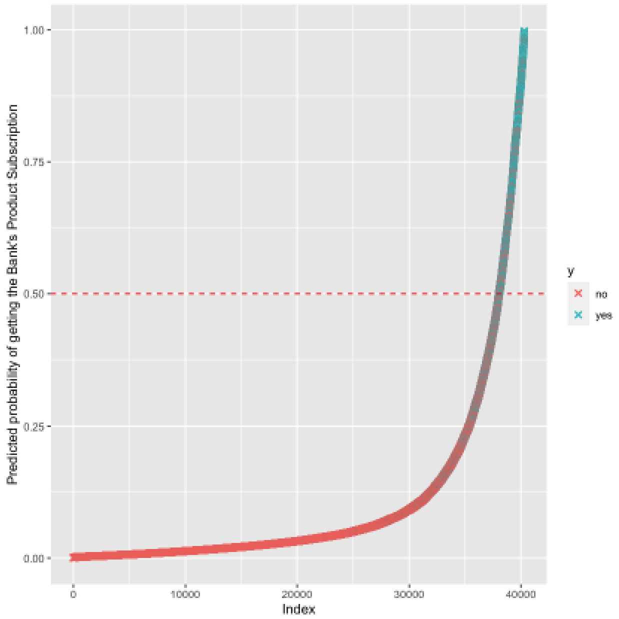
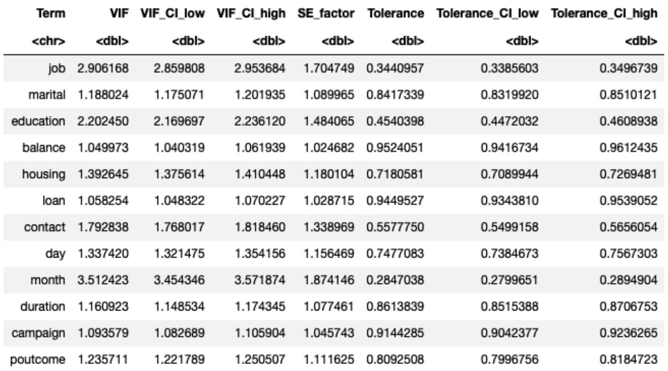


Fig. 12: Model 3 Predictions



Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
job	2.906168	2.859808	2.953684	1.704749	0.3440957	0.3385603	0.3496739
marital	1.188024	1.175071	1.201935	1.089965	0.8417339	0.8319920	0.8510121
education	2.202450	2.169697	2.236120	1.484065	0.4540398	0.4472032	0.4608938
balance	1.049973	1.040319	1.061939	1.024682	0.9524051	0.9416734	0.9612435
housing	1.392645	1.375614	1.410448	1.180104	0.7180581	0.7089944	0.7269481
loan	1.058254	1.048322	1.070227	1.028715	0.9449527	0.9343810	0.9539052
contact	1.792838	1.768017	1.818460	1.338969	0.5577750	0.5499158	0.5656054
day	1.337420	1.321475	1.354156	1.156469	0.7477083	0.7384673	0.7567303
month	3.512423	3.454346	3.571874	1.874146	0.2847038	0.2799651	0.2894904
duration	1.160923	1.148534	1.174345	1.077461	0.8613839	0.8515388	0.8706753
campaign	1.093579	1.082689	1.105904	1.045743	0.9144285	0.9042377	0.9236265
poutcome	1.235711	1.221789	1.250507	1.111625	0.8092508	0.7996756	0.8184723

Fig. 13: VIF Scores

II. Time Series Analysis

Analysing a sequence of data points which have been collected over a certain period of time is called time series analysis. In time series analysis the main objective is to statistically analyse and understand how different variables change over time. Different statistical models helps us to fetch the actual nature of a time series data and forecast the future trend based on the collected sample data. In this study our prime objective is to estimate and evaluate a suitable model for the following data.

A. Data Description

This analysis has been done on the *Flight Departure* data which is basically a monthly time series of average number of departures from Ireland via airports, commencing in 2010

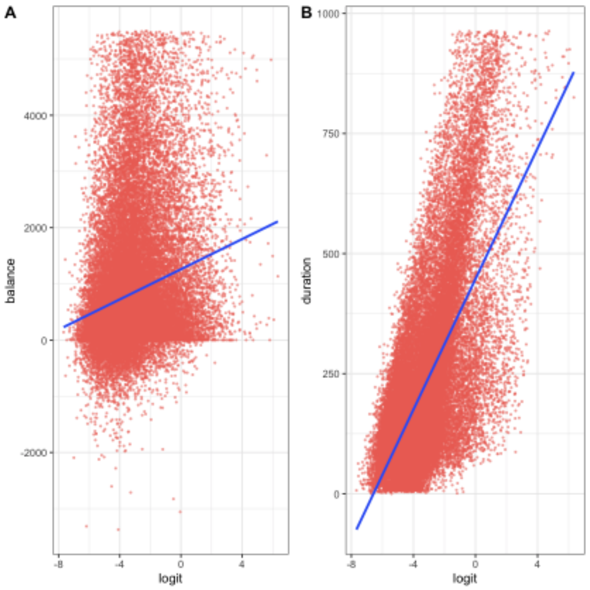


Fig. 14: Continuous Variables vs Log (odds)

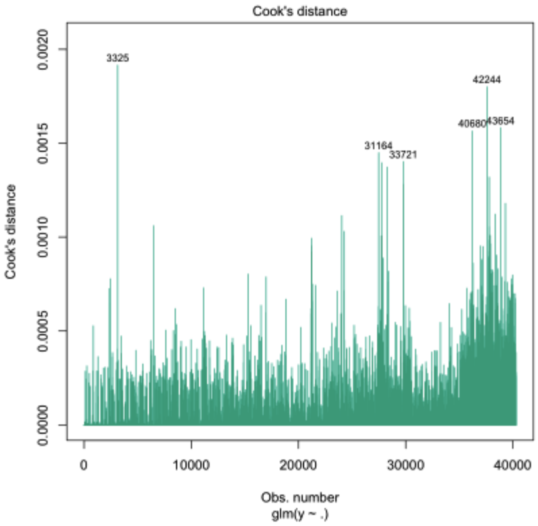


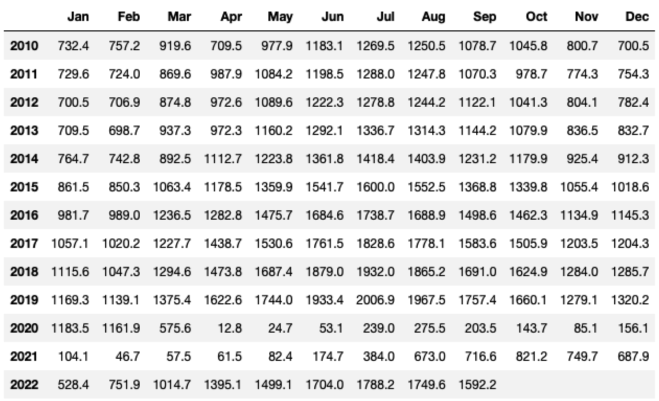
Fig. 15: Cook’s Distance

to September 2022. Initially the data file was consisted of a data-frame with 2 variables with 153 observations but before starting our analysis, we have converted the data-frame into a time series data and the Figure 16 gives a glimpse of that.

B. Objective

Firstly, A preliminary assessment of the nature and components of the raw time series, using appropriate visualisations.

Secondly, Estimating and discussing different suitable time series models namely, Exponential Smoothing, ARIMA, SARIMA or other simple time series models using appropriate diagnostic tests and checks.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010	732.4	757.2	919.6	709.5	977.9	1183.1	1269.5	1250.5	1078.7	1045.8	800.7	700.5
2011	729.6	724.0	869.6	987.9	1084.2	1198.5	1288.0	1247.8	1070.3	978.7	774.3	754.3
2012	700.5	706.9	874.8	972.6	1089.6	1222.3	1278.8	1244.2	1122.1	1041.3	804.1	782.4
2013	709.5	698.7	937.3	972.3	1160.2	1292.1	1336.7	1314.3	1144.2	1079.9	836.5	832.7
2014	764.7	742.8	892.5	1112.7	1223.8	1361.8	1418.4	1403.9	1231.2	1179.9	925.4	912.3
2015	861.5	850.3	1063.4	1178.5	1359.9	1541.7	1600.0	1552.5	1368.8	1339.8	1055.4	1018.6
2016	981.7	989.0	1236.5	1282.8	1475.7	1684.6	1738.7	1688.9	1498.6	1462.3	1134.9	1145.3
2017	1057.1	1020.2	1227.7	1438.7	1530.6	1761.5	1828.6	1778.1	1583.6	1505.9	1203.5	1204.3
2018	1115.6	1047.3	1294.6	1473.8	1687.4	1879.0	1932.0	1865.2	1691.0	1624.9	1284.0	1285.7
2019	1169.3	1139.1	1375.4	1622.6	1744.0	1933.4	2006.9	1967.5	1757.4	1660.1	1279.1	1320.2
2020	1183.5	1161.9	575.6	12.8	24.7	53.1	239.0	275.5	203.5	143.7	85.1	156.1
2021	104.1	46.7	57.5	61.5	82.4	174.7	384.0	673.0	716.6	821.2	749.7	687.9
2022	528.4	751.9	1014.7	1395.1	1499.1	1704.0	1788.2	1749.6	1592.2			

Fig. 16: Flight Departure Data

Finally, Forecasting the number of departures in the first 6 months of 2021 and discussing the choice of an ‘optimum’ model for this series.

C. Assessment of the raw Time Series

Figure 17 depicts a clear seasonality in the graph. There is a dramatic drop in the flight departure in between the year 2020 to 2021 which justifies the COVID-19 pandemic locked down situations. The horizontal line with violet color indicates the mean of the number of departures in the data-set and it is clear that the mean of the time series data is not constant over the time which suggests that the given data is non-stationary.

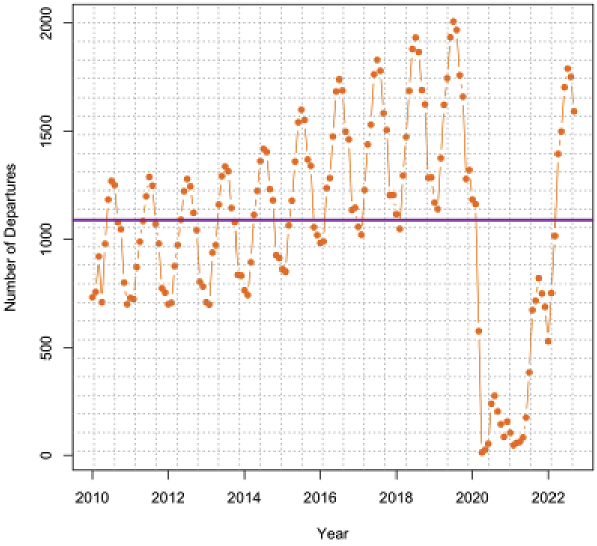


Fig. 17: Time Series Plot

The box-plot per year (in figure 18) gives us a summary of the trend over the years which looks like an envelope for the data. For this chart, we have plot number of departures against the floor value of the times of the data-set. In R, ’time()’ function returns a floating point variable that interpolates the time interval between the years. So we have used floor values to cuts it down to the actual year. Again the drastic fall in the number of flight departures justifies the reason explained before.

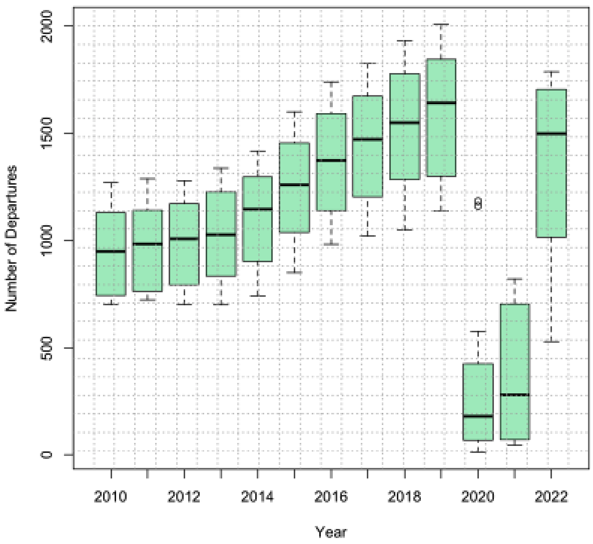


Fig. 18: Box-plot per year

We have plot number of departures for each individual years starting from 2010 to 2022 (see Figure 19) and the drastic changes in the years 2020 and 2021 is clearly visible in this.

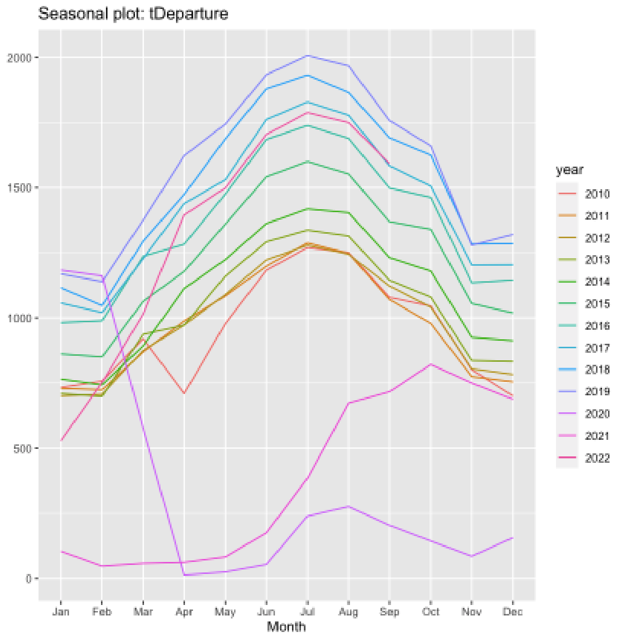
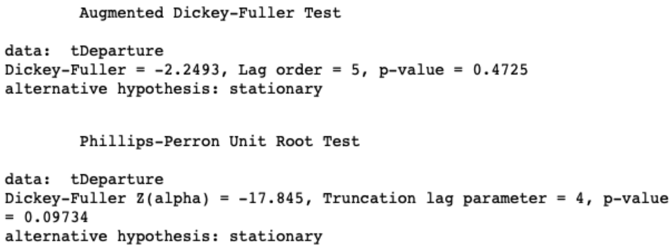


Fig. 19: Seasonal Plot

D. General Diagnostics of the Time Series

In this section we have tried to give appropriate diagnostic tests and graphical visualisation for different component of the time series data.

1) Test for Stationarity: Augmented Dickey-Fuller (ADF) test is commonly used to test whether the time series data is stationary or not. ADF test assumes the null hypothesis as the data is non-stationary and hence a significantly larger p-value suggests the alternative hypothesis. In Figure 20 the P-value for ADF test is 0.4725 which indicates that there is no such statistical evidence to reject the null hypothesis and hence our data is non-stationary. The *Phillips-Perron* (PP) test works in the similar way.



```
        Augmented Dickey-Fuller Test

data:  tDeparture
Dickey-Fuller = -2.2493, Lag order = 5, p-value = 0.4725
alternative hypothesis: stationary


        Phillips-Perron Unit Root Test

data:  tDeparture
Dickey-Fuller Z(alpha) = -17.845, Truncation lag parameter = 4, p-value
= 0.09734
alternative hypothesis: stationary
```

Fig. 20: ADF and PP test

In Figure 21, we have tried to plot the time series data and to smooth out the seasonal variation we have applied moving average with order 5 (in red colour) and further moving average of order 20. The yellow straight line indicates the mean of the data. Hence we now have a clear vision of upward trend (before year 2020) in the series.

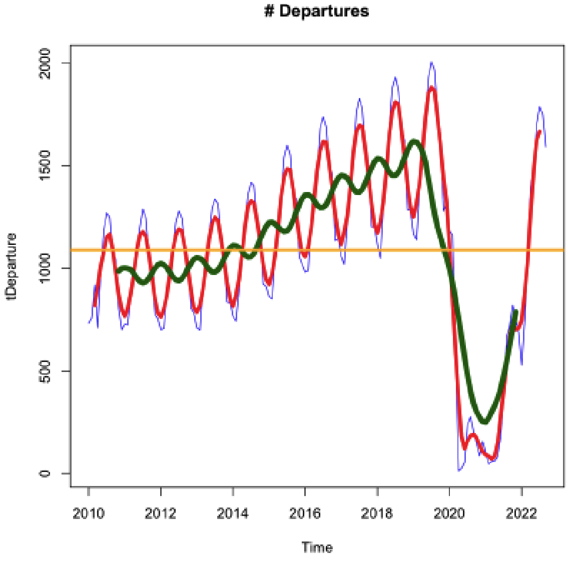


Fig. 21: Visualising Trend

2) Making the data stationary: Differencing is a method which can help to stabilize the mean of a time series data. By removing changes in different levels of a time series, it helps to eliminate or reduce the trend and seasonality present in it. Here we have taken first order differncing which successfully removed the upward trend from the data (see Figure 22). We have also performed ADF test after first order differencing and got the p-value smaller than 0.01 which suggests that the data has become stationary.

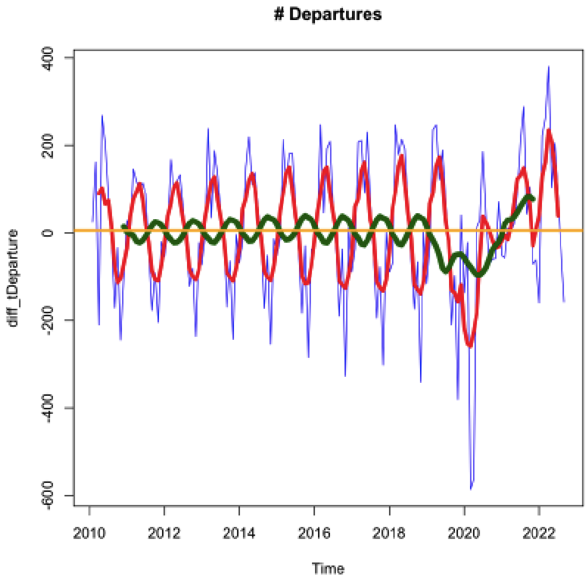


Fig. 22: Data after First Differencing

3) Homoscedasticity Test: ‘The Box-Cox linearity plot is a plot of the correlation between Y and the transformed X for given values of λ. That is, λ is the coordinate for the horizontal axis variable and the value of the correlation between Y and the transformed X is the coordinate for the vertical axis of the plot. The value of λ corresponding to the maximum correlation (or minimum for negative correlation) on the plot is then the optimal choice for λ’ [3]. Since in our case, the value of λ in the Box-Cox test is nearly 1 (1.22 exactly), we can interpret that the variability in the time series data is nearly constant. Hence no transformation is need to make the variance constant.

4) Identifying and Removing Seasonality using ACF and PACF: Auto Correlation Function (ACF) measures the correlation of any series with its lagged values and the Partial Auto Correlation Function (PACF) measures the correlation of any series with its lagged values but also incorporates the correlations among the different lagged values in the series.

Figure 23 shows the ACF and PACF plot after the first diffenencing of the time seris data. Here, a repetitive pattern in every 6 lags can be clearly observed in both the chart which suggests that though the data is stationary but the seasonality factor is still present after performing first order differencing.

To remove seasonality, we have taken differencing with lag order 12 since from the above discussion it is clear that there is 12 months seasonality present in this data. In Figure 24, we cannot see the previous repetitive patterns in the ACF and the PACF plots and hence we can conclude that now the data is stationary and there is no seasonal trend present.

E. Models Selection

Based on the above diagnostic tests of our time series data, we have chosen the following models and done a comparison

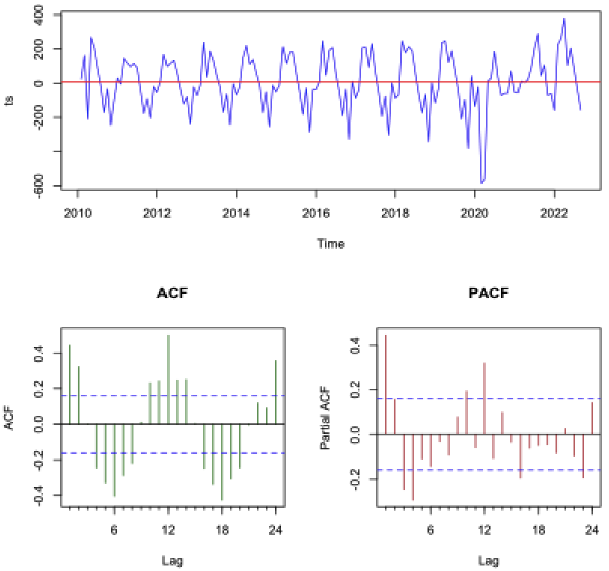


Fig. 23: ACF and PACF after First Differencing

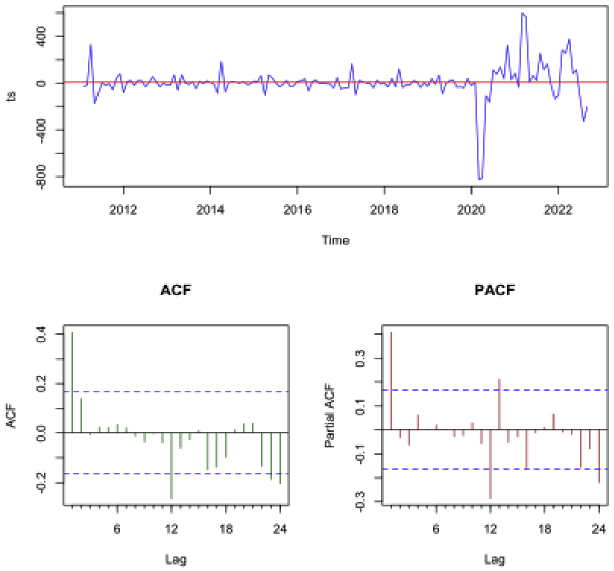


Fig. 24: ACF and PACF after removing Seasonality

study to find out the best fitted model among others by giving the analytical justifications for rejection intermediate models.

1) ARIMA or SARIMA Model: After removing trend and seasonality from the data, we can observe that there are significant spikes in lag 12, lag 24 in the ACF and PACF graph which suggests that a Seasonal ARIMA model with non-seasonal orders (p,d,q) and the seasonal orders (P,D,Q) would be a appropriate here. Now we need to find out best non-seasonal and seasonal orders for sarima model and for that we have tried a several models with different orders and done a comparative study.

- *Model 1* - From the PACF graph (in Figure 24), there is a significant spike at lag 1 where as in ACF graph there are two significant spikes at lag 1 and 2. We should always start choosing the model orders from minimum possible vales to avoid the model complexity. So we have chosen the non-seasonal orders for the sarima model as (p,d,q) as (1,1,0). Since we have done first order differencing to make the data stationary, we have taken d as 1. Now for seasonal orders we have chosen (P,D,Q) as (2,1,0). Here we have taked D as 1 since we have applied 1 lag differencing to remove seasonality.
- *Model 1 Summary* - In the Figure 25 ACF chart of residuals, there is no significant spikes i.e all the residuals are within the significance limit and hence the residuals appear to be white noise. The *Box-Ljung* test assumes the null hypothesis that the auto correlation of the residuals are zero. Here for this model the large p-value (0.55 approx) fails to reject the null hypothesis (at 5% significance level) suggesting that there is no statistically significant evidence of the presence of auto-correlation among the residuals.
- *Models Comparison and conclusion* - Like model 1, we have tried to fit other sarima models with different logical choices of non-seasonal and seasonal orders. The table II compares all the models on the basis of AIC scores and number of spikes in the ACF for the residuals and suggests that among all the models, model 1 has lowest AIC score and there is no spikes in the AFC chart and it satisfies the Box-Ljung test. Hence, the model 1 would be the best choice to forcast by all means.

TABLE II: SARIMA Models Comaparison Table

sl.	Model Name	AIC	ACF spikes
1	ARIMA(1,1,0)(2,1,0)[12]	1735.81	0
2	ARIMA(0,1,1)(2,1,0)[12]	1739.40	0
3	ARIMA(2,1,0)(0,1,2)[12]	1735.93	1
4	ARIMA(2,0,0)(2,1,0)[12]	1745.71	0
5	ARIMA(2,1,0)(1,1,0)[12]	1763.74	1

2) Exponential Smoothing: In the section ‘General Diagnostics of the Time Series Data’ we have seen that trend component and seasonal component are present in our data. So for exponential smoothing method, we should go with ‘Holt-Winter Seasonal Method’ as it is able to deal with trend and the seasonal variation simultaneously. Now in this method, we need to check whether the the additive seasonal smoothing or the multiplicative seasonal smoothing would be best for our time series data.

- *Models Comparison and Conclusion* - In the figure 26, we have compared four different Holt-Winter Seasonal Methods and found that only the model 4 i.e. Holt-Winter Multiplicative Damped model satisfies the Box-Ljung test which suggests that there is no auto-correlation among the residuals. The RMSE value for this model is 106.90 which is comparatively low as well.

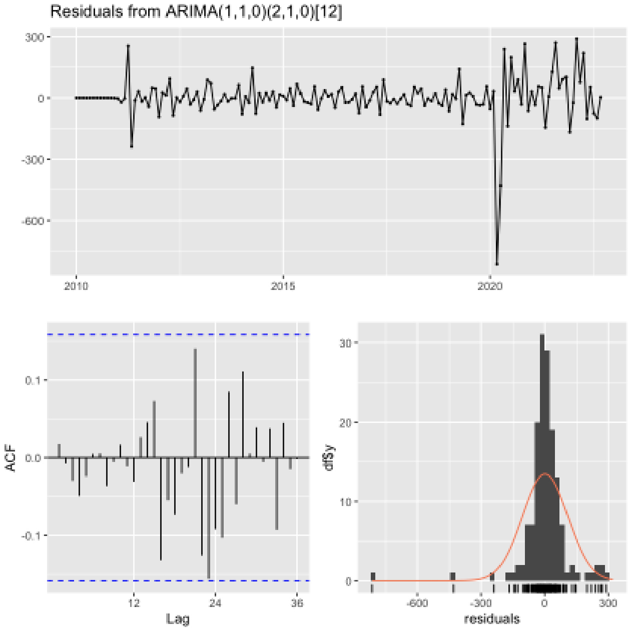
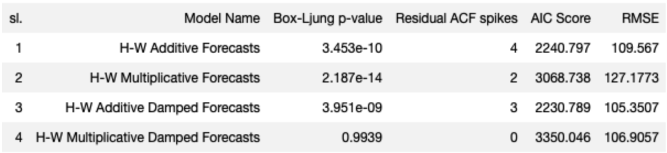


Fig. 25: Model 1 Residual Plot



sl.	Model Name	Box-Ljung p-value	Residual ACF spikes	AIC Score	RMSE
1	H-W Additive Forecasts	3.453e-10	4	2240.797	109.567
2	H-W Multiplicative Forecasts	2.187e-14	2	3068.738	127.1773
3	H-W Additive Damped Forecasts	3.951e-09	3	2230.789	105.3507
4	H-W Multiplicative Damped Forecasts	0.9939	0	3350.046	106.9057

Fig. 26: Holt-Winter Models Comparison

3) Simple Time Series Models: We have also tried to fit simple time series models like linear regression model and logarithmic regression model but which are not useful in our given data. Figure 27 clearly depicts that both the models have failed to capture the trend and seasonality in the series and hence we can not use these models for forecasting.

F. Forecasting

We have forecast the number of departures in first 6 months of 2021 and compared that forecast with the original figures and finally decided the optimum model for the given time series data.

1) Forecasting with SARIMA model: Figure 28 shows the forecast values of first 6 months of 2021 by SARIMA model with 80% and 90% confidence intervals. Here the red line corresponds to the observed values and the blue line corresponds to the forecast values. It is clear that the forecast value by the SARIMA model captures maximum variability of the original series.

Figure 29 reflects the model accuracy in forecasting. The root mean square error (RMSE) and the Mean Absolute Error

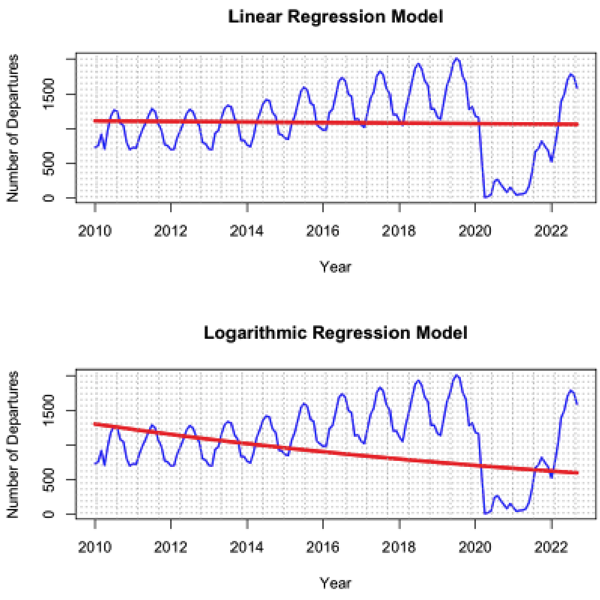


Fig. 27: Simple Time Series Models

```
         Point Forecast       Lo 80     Hi 80     Lo 95      Hi 95
Jan 2021      66.810679  -74.42505 208.0464 -149.1907   282.8121
Feb 2021      35.297427 -208.69292 279.2878 -337.8536   408.4485
Mar 2021      26.281091 -303.95847 356.5207 -478.7767   531.3389
Apr 2021       1.599145 -402.09507 405.2934 -615.7979   618.9962
May 2021     122.777383 -344.90312 590.4579 -592.4782   838.0330
Jun 2021     263.680375 -260.95675 788.3175 -538.6829  1066.0436
```

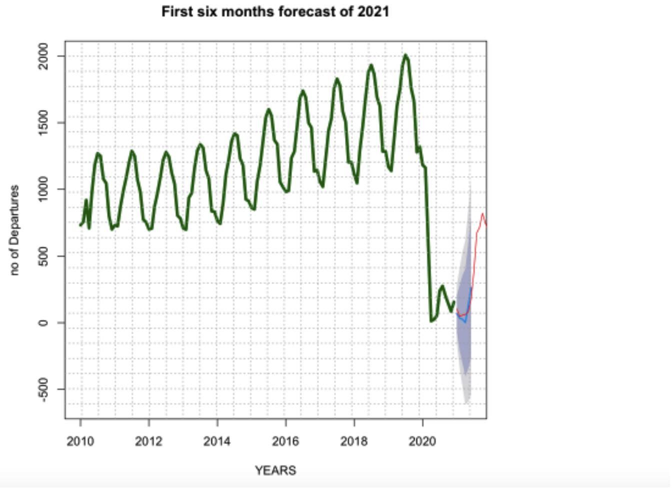
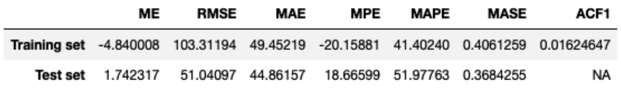


Fig. 28: Forecast with SARIMA Model

(MAE) values are respectively 51.04 and 44.86 in forecasting using SARIMA model.



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-4.840008	103.31194	49.45219	-20.15881	41.40240	0.4061259	0.01624647
Test set	1.742317	51.04097	44.86157	18.66599	51.97763	0.3684255	NA

Fig. 29: SARIMA Forecast Accuracy

2) Forecasting with H-W Damped Multiplicative Model Model: Figure 30 shows the forecast values of first 6 months of 2021 by Holt-Winter model with 80% and 90% confidence intervals. Like before the red line corresponds to the observed values and the blue line corresponds to the forecast values. It is clear that the forecast value by the Holt-Winter model fails to captures maximum variability of the original series.

```
         Point Forecast       Lo 80    Hi 80       Lo 95    Hi 95
Jan 2021       142.5787   48.436092 236.7212   -1.399972 286.5573
Feb 2021       139.3749   -1.882629 280.6325  -76.659864 355.4097
Mar 2021       153.1588  -54.007504 360.3252 -163.674747 469.9924
Apr 2021       160.1495 -113.480522 433.7794 -258.331483 578.6304
May 2021       176.8259 -194.091564 547.7433 -390.443389 744.0951
Jun 2021       195.0967 -298.993905 689.1872 -560.549657 950.7430
```

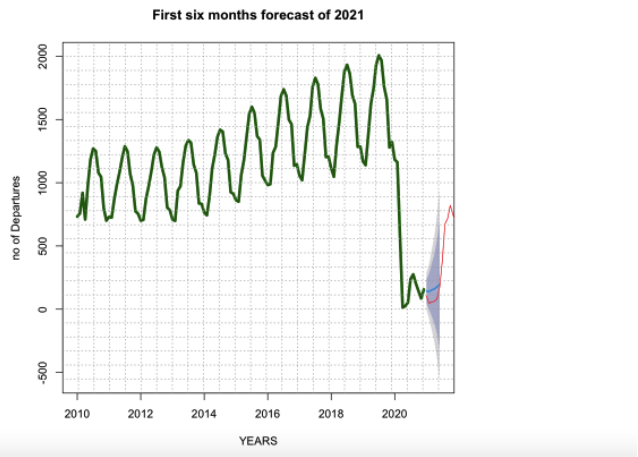
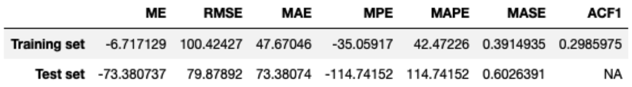


Fig. 30: Forecast with Holt-Winter Model

Figure 29 reflects the model accuracy in forecasting. The root mean square error (RMSE) and the Mean Absolute Error (MAE) values are respectively 80 (nearly) and 73.38 in forecasting using Holt-Winter model.



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-6.717129	100.42427	47.67046	-35.05917	42.47226	0.3914935	0.2985975
Test set	-73.380737	79.87892	73.38074	-114.74152	114.74152	0.6026391	NA

Fig. 31: Holt-Winter Forecast Accuracy

G. Conclusion

In the above through discussion starting from model selection to forecast, it is clear that the Seasonal ARIMA model is appropriate for the given time series data. More specifically, the arima model with the non-seasonal orders (p,d,q) as (1,1,0) and the seasonal orders (P,D,Q) as (2,1,0) with seasonality period 12 has performed best among all the other models discussed above.

References

[1] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer, 2004, vol. 26.

[2] *R squared in logistic regression*, http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/.

[3] *NIST Information Technology Laboratory*, https://www.itl.nist.gov/div898/handbook/eda/section3/boxcoxli.html.