A Time Series Analysis of Departures from Ireland Airport and A Bank Product Sales Prediction Using Logistic Regression Model

Jia Lin
Student Number: 22117644
Postgraduate Diploma in Science in Data Analytics
National College of Ireland
x22117644@student.ncirl.ie

I. Introduction

This paper presents statistical forecasts on two datasets. Before the specific analysis, the provided datasets haven been cleaned and divided into training and test sample sets.

The first topic is a prediction of the number of departures from Ireland via airport, by selecting a Time Series Analysis (TSA) model. There are several classes of TSA models, such as Exponential Smoothing (EST) models, Auto-Regressive Integrated Moving Average (ARIMA) models and some other simple models. The first objective of this paper is to identify an optimum model for the provided dataset, which contains data of departures from Ireland airport. The process of determination of the most appropriate forecasting model has been discussed. For models in the same class, the value of Akaike's Information Criterion (AIC) has been set to be the determination factor. For models of diverse classes, the lowest value of Root Mean Squared Error (RMSE) is expected.

The second analysis is used to estimate a bank product sales based on details of marketing compaign data. A Logistical Regression (LR) model has been built to complete the forecast and a dimension-reduction technique, called Principle Components Analysis (PDA), is employed to turn down the complexity of the specific model. The Pseudo R^2 test and ANOVA test are used to compare different models then select an optimal binary LR model. Finally, a confusion matrix have been created to examine the predicted probabilities and classifications of the created LR model and complete the cross-validation on a test dataset.

II. Related Work

Some previous studies has been done to using the holt-winters exponential smoothing method of additive models to predict the departures from Hasanudin Airport based on the seasonal time series from 2009 to 2019 [1], their dataset is similar with the sample dataset in this paper, the approach taken the EST model also selected the additive approach. Another research created SARIMA Time Series Model to forecast the passenger flow for airport terminal [2], however, their work concentrated on analysing the number of passengers in the airport in different time period in a day, based on the sample set during a whole year. And they compared many ARIMA models with different orders of parameters of auto-regression and moving average models. Some other work implemented the prediction of departures from airport using Machine Learning [3] and Deep Learning Algorithms [4].

Related work has been done by building a logistic regression model to prediction of usage of e-Banking [5], in the evaluation section they use Hosmer and Lemeshow test to present the goodness-of-fit of their model. Another work identified the "microeconomic" predictors to forecast the bank failure or defect by creating a logistic regression model [6]. By comparing "four data mining models: logistic regression, decision trees, neural network and support vector machine", the research [7], predicted the success of telemarketing calls to sell a bank product.

III. Time Series Analysis

A. Dataset Description and Data Pre-processing

A dataset in CSV format is provided to do the time series analysis. The dataset presents the number of departures from Ireland airport in a time series from Jan. 2010 to Sep. 2022. The aim of this paper on this topic is to select an appropriate time series forecast model to predict the number of departures in the future. This dataset is composed of 153 rows or observations, and 2 columns, one is the "months", the other is the "departures" which indicates the number of departures. The variable *departures* is the dependent variable, and the variable *months* is the independent variable.

Before conducting any data analytics, the data should be cleaned and ready for training. Normally, the data of type String, Numeric, in particular, the data of type Date are considered. Values in two columns in the original table are type of String. Values of months variable has been converted to Date format and values of departures variable has been transformed into Numeric type.

There is no NULL value or error value in the data set, the first six observations has presented as an example in Fig. 1.

```
> head(df_departures)
      Months Departures
1 2010-01-01      732.4
2 2010-02-01      757.2
3 2010-03-01      919.6
4 2010-04-01      709.5
5 2010-05-01      977.9
6 2010-06-01     1183.1
```

Fig. 1: The Dataframe of Departures

B. Initial Time Series Analysis

A raw time series plot has been made for this data sample, see Fig. 2. By observing this visualisation of the original

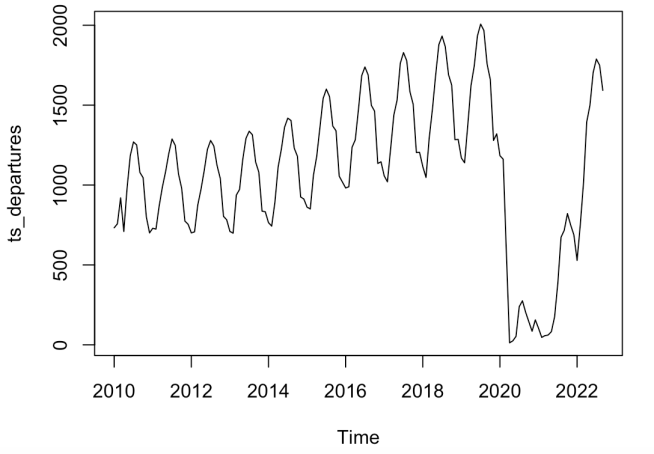

Fig. 2: A Raw Time Series Plot of Departures

dataset, the time series object based on such data sample (called ts_departures) is not a stationary. The trend and seasonal pattern can be seen from 2010 to 2020, then a drop down happened at the beginning of 2020. These features have been shown in the monthly plot (see Fig. 3) and the seasonal plot (see Fig. 4) of ts_departures.

In order to illustrate the characters of this ts_departures time series more clearly, this object has been decomposed into three components, seasonal, trend and irregular components, by using STL approach, in which STL stands for Seasonal and Trend decomposition using Loess. The benefit of using such STL approach is avoiding shortcomings of classic decomposition additive and multiplicative models. For instance, the seasonal pattern is supposed to repeat every year does not suit this sample set. The graph result presents the original data, the seasonal plot, the trend of such time series and the irregular components (see Fig. 5). In particular, this STL model has been applied to the logarithmic transformation

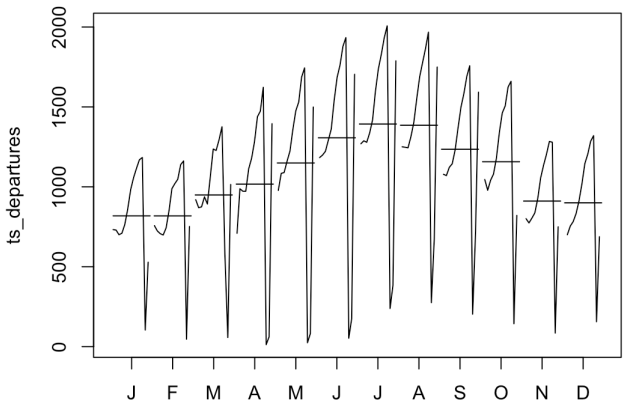

Fig. 3: A Monthly Plot of Departures

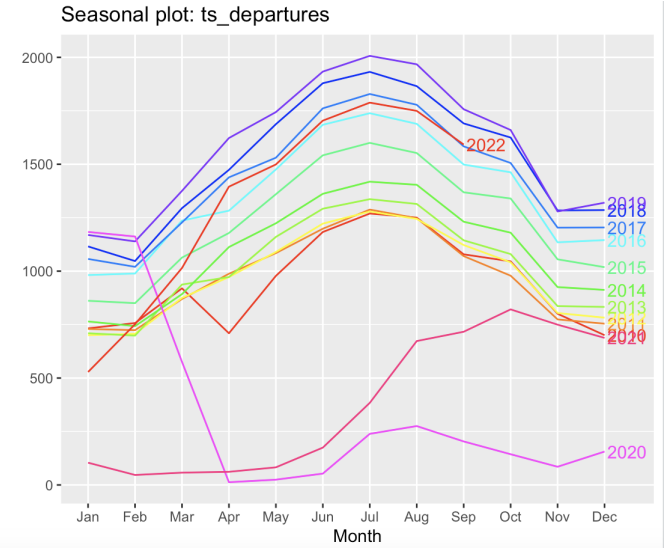

Fig. 4: A Seasonal Plot of Departures

of ts_departures. This logarithm form, named log_departures, strengthens the features of such time series object.

IV. Time Series Forecasting Models Selection

Before choosing a optimal forecasting method, the data sample has been divided into a training dataset and a test dataset. The training set contains data from the beginning of 2010 to the end of 2020, and the test set is composed of data from Jan. 2021 to Sep. 2022. The Fig. 6 shows a raw plot of data in the training set. Such training and test sets approach has a faster performance than the cross-validation way.

There are many forecasting models to predict the future data of a time series object. For example, the ETS (Exponential Smoothing) models, the ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal ARIMA) models and some simple forecasting methods, such as Average method, Random Walk method, Seasonal naïve method, Drift method.

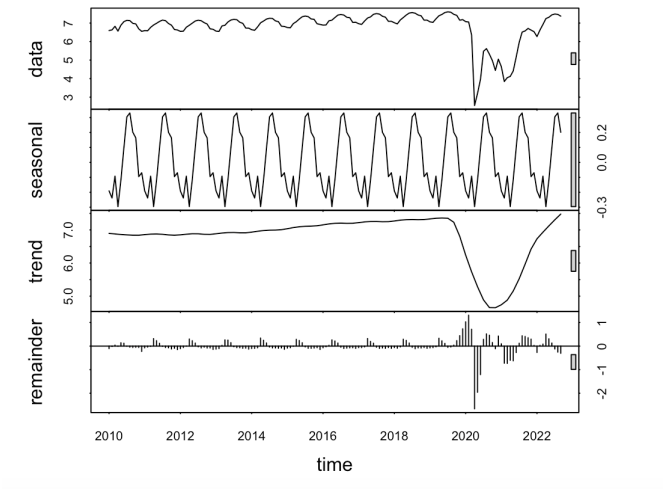

Fig. 5: The Decomposition of Departures using STL Method

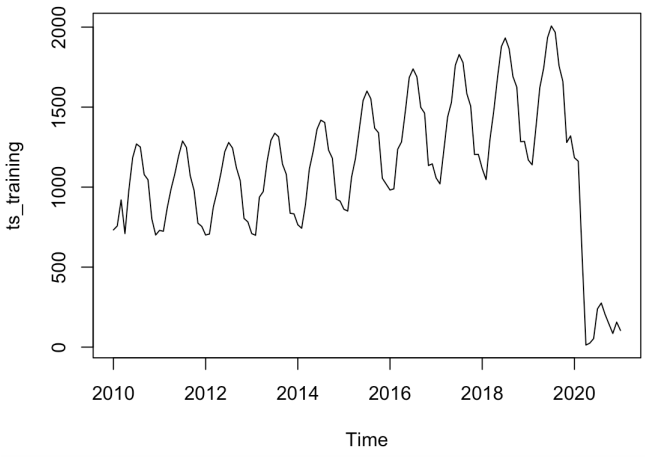

Fig. 6: A Time Series Plot of Training Set of Departures

In this paper, a optimum forecasting model is determined by identifying a model with lowest RMSE (Root Mean Squared Error), and for models in the same class, the smallest AIC (Akaike's Information Criterion) is the decision factor.

A. Exponential Smoothing Models

The ETS models are one of the most popular forecasting models for time series objects. Such class of models predicts the future value of a time series dataset by setting different weights to the past observations. The latest observation has the highest weight, when the observations getting older, the weights are broken down exponentially. There are three smoothing parameters of ETS models, level which is the local mean value of the time series, the trend and the seasonal component. In terms of the number of smoothing parameters, there are three classic ETS models:

- A simple exponential model, also called single exponential model, fits a time series with a level and an irregular component or error term at a time.
- A double exponential model, named Holt exponential model, fits a time series with a level and a trend.
- A triple exponential model, or Holt-Winters exponential model, fits a time series with a level, a trend and a seasonal components.

Since the decomposition approaches comprise additive and multiplicative models, there are many combinations of the smoothing parameters when applying the ETS models. Taking the classic ETS models for example, see Table I, R provides functions for ETS models. In the ets() function, ts indicates a specific time series, the value of model presents the combination of three capital letters. These letters denote the level, trend and seasonal type respectively, in which A represents the additive, N represents the none which indicates a smoothing parameter is not considered in a corresponding model. In addition, more options are provided, M for multiplicative, Z for automatically.

In order to identify a most appropriate ETS model, a function $\mathsf{ets}(\mathsf{ts}, \mathsf{model} = \text{“ZZZ”})$ has been used to select the ETS model automatically, see Fig. 7. The $\mathsf{ETS}(\mathsf{A}, \mathsf{Ad}, \mathsf{A})$ model has been chosen and this is the Holt-Winters method with a damped trend [8] and additive seaonality. In which, Ad is a damped variant of a additive trend term, the corresponding damped variant of a multiplicative trend component is Md.

TABLE I: Classic ETS Models

Type	Parameters Fit	Functions in R
Single	level	$\mathsf{ets}(\mathsf{ts}, \mathsf{model} = \text{“ANN”})$ $\mathsf{ses}(\mathsf{ts})$
Holt	level, trend	$\mathsf{ets}(\mathsf{ts}, \mathsf{model} = \text{“AAN”})$ $\mathsf{holt}(\mathsf{ts})$
Holt-Winters	level, trend, seasonal	$\mathsf{ets}(\mathsf{ts}, \mathsf{model} = \text{“AAA”})$ $\mathsf{hw}(\mathsf{ts})$

R.J. Hyndman and G. Athanasopoulos, in their book "Forecasting: principles and practice" [9], introduce formulas of the Holt-Winters' additive, multiplicative and damped method. The following equation of the $\mathsf{ETS}(\mathsf{A}, \mathsf{Ad}, \mathsf{A})$ model, is inferred from the damped method with the multiplicative seasonality, by replacing the multiplicative seasonality with the additive seasonality.

$$\begin{aligned}
\hat{y}_{t+h|t} &= [l_t + (\phi + \phi^2 + \ldots + \phi^h) b_t] s_{t+h-m(k+1)} \\
l_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + \phi b_{t-1}) \\
b_t &= \beta^*(l_t - l_{t-1}) + (1-\beta^*)\phi b_{t-1} \\
s_t &= \gamma(y_t - l_{t-1} - \phi b_{t-1}) + (1-\gamma) s_{t-m}
\end{aligned}$$

In which, m is the number of observations per year, for this monthly time series $m = 12$, and "k is the integer part of $(h-1)/m$ which ensures that the estimates of the seasonal indices used for forecasting come from the final year of the sample. " [9]

The accuracy of such model is displayed below, see Fig. 8, the RMSE is 100.58. By applying the forecast() function

```
> fit_ets <- ets(ts_training, model="ZZZ")
> fit_ets
ETS(A,Ad,A)

Call:
 ets(y = ts_training, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.9999
    beta  = 0.1413
    gamma = 1e-04
    phi   = 0.8

  Initial states:
    l = 962.4632
    b = 4.7834
    s = -208.2974 -202.7917 72.2755 130.9762 307.6094 349.7194
          233.3971 73.3278 -65.9956 -144.2698 -278.4757 -267.4752

  sigma:  107.7625

     AIC     AICc      BIC
1897.832 1903.886 1949.723
```

Fig. 7: The Automated ETS Model

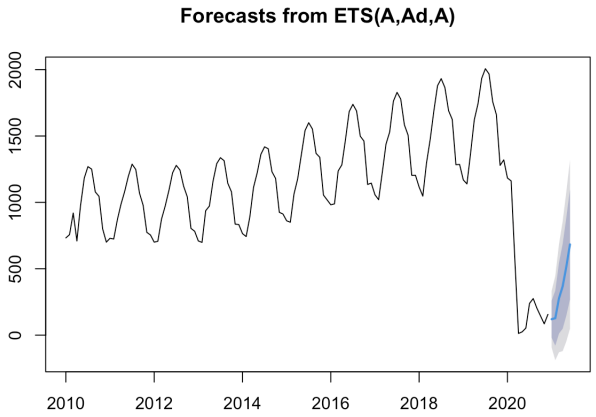

Fig. 10: The Forecast from ETS Model(A, Ad, A)

to this model, the departures of the first 6 months of 2021 is predicted, see Fig. 9 and the corresponding plot of this forecasts has been presented in Fig. 10.

```
> round(accuracy(fit_ets),2)
                    ME   RMSE   MAE    MPE  MAPE MASE ACF1
Training set -2.43 100.58 52.82 -30.07 41.14 0.27 0.28
```

Fig. 8: The Accuracy of ETS(A, Ad, A) Model

```
> fc_fit_ets <- forecast(fit_ets, 6)
> fc_fit_ets
         Point Forecast     Lo 80     Hi 80      Lo 95     Hi 95
Jan 2021       120.1400 -17.96313  258.2432  -91.07053  331.3506
Feb 2021       127.7165 -78.91624  334.3493 -188.30103  443.7341
Mar 2021       276.7834  11.60873  541.9580 -128.76624  682.3329
Apr 2021       366.9147  48.54473  685.2846 -119.99014  853.8195
May 2021       515.7952 147.98794  883.6024  -46.71744 1078.3078
Jun 2021       683.4642 269.21589 1097.7125   49.92609 1317.0023
```

Fig. 9: The Forecast of Departures of the First 6 Months of 2021 by using ETS(A, Ad, A) Model.

By comparing the value of AIC among the selected model (Fig. 7), single exponential model (Fig. 11) and double exponential model (Fig. 12), the selected model has the smallest AIC value, indicates it is the most appropriate model for this training dataset.

Additionally, the plot of comparison among the original time series, the Holt-Winters additive and multiplicative model, illustrates that the additive model is fit this datasets more, see Fig. 13. Although the multiplicative model has a lower value of RMSE, approximate $100.42 < 100.58$. That is the reason why the automated ETS model suggest the damped trend with the additive seasonality model.

B. ARIMA/SARIMA Models

A stationary time series has stable properties, such as a constant mean and variance. And there is no trend and

```
> fit_single <- ets(ts_training, model = "ANN")
> fit_single
ETS(A,N,N)

Call:
 ets(y = ts_training, model = "ANN")

  Smoothing parameters:
    alpha = 0.9999

  Initial states:
    l = 762.1383

  sigma:  160.1681

     AIC     AICc      BIC
1988.638 1988.825 1997.286
```

Fig. 11: The Single Exponential Model

```
> fit_holt <- ets(ts_training, model = "AAN")
> fit_holt
ETS(A,Ad,N)

Call:
 ets(y = ts_training, model = "AAN")

  Smoothing parameters:
    alpha = 0.8148
    beta  = 0.8148
    phi   = 0.8

  Initial states:
    l = 764.9002
    b = 56.2126

  sigma:  147.4473

     AIC     AICc      BIC
1969.709 1970.381 1987.006
```

Fig. 12: The Double Exponential Model

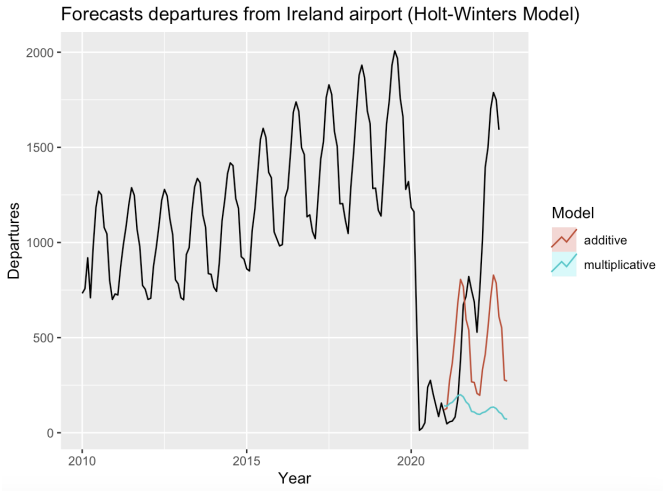

Fig. 13: The comparison between Holt-Winters Additive and Multiplicative Models

seasonality of a stationary time series, which makes it rubost to predict future values based on it.

ARIMA, stands for autoregressive integrated moving average, and SARIMA, stands for Seasonal ARIMA, are statistical forecasting models, which are designed for stationary time series objects, or any time series can be transformed to a stationary time series.

The Augmented Dickey-Fuller (ADF) Test is used to indicate weather a time series is a stationary or not. By applying the function adf.test() in packge tseries to ts_training dataset, the result, p-value is not significant $0.7658 > 0.05$, shows that it is not a stationary time series, see Fig. 14.

```
> adf.test(ts_training)

        Augmented Dickey-Fuller Test

data:  ts_training
Dickey-Fuller = -1.5447, Lag order = 5, p-value = 0.7658
alternative hypothesis: stationary
```

Fig. 14: The Dickey-Fuller Test on ts_training Time Series

Differencing is an approach to calculate the difference between successive observations, and it is used to making a non-stationary time series to a stationary time series by removing trend and seasonality. By applying 1 or 2 times differencing to a non-stationary time series, it can be made as a stationary time series. The function $\mathsf{diff}(\mathsf{ts}, \mathsf{differences} = \mathsf{d})$ is used to implement the differencing procedure, in which ts is the specific time series and the value d is the frequency of differencing. The function ndiffs() is used to determine the best value of the value d. The result of $\mathsf{ndiffs}(\mathsf{ts\_training})$ shows $\mathsf{d} = 1$. A new time series, stat_training, is created by executing the function $\mathsf{diff}(\mathsf{ts\_training}, \mathsf{differences} = 1)$.

By applying the ADF test to the stat_training time series, p-value is significant indicates the stat_training is a stationary time series (Fig. 15). The mean-variance stationary time series, stat_training is presented in Fig. 16.

```
> adf.test(stat_training)

        Augmented Dickey-Fuller Test

data:  stat_training
Dickey-Fuller = -7.2297, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(stat_training) : p-value smaller than printed p-value
```

Fig. 15: The Dickey-Fuller Test on stat_training Time Series

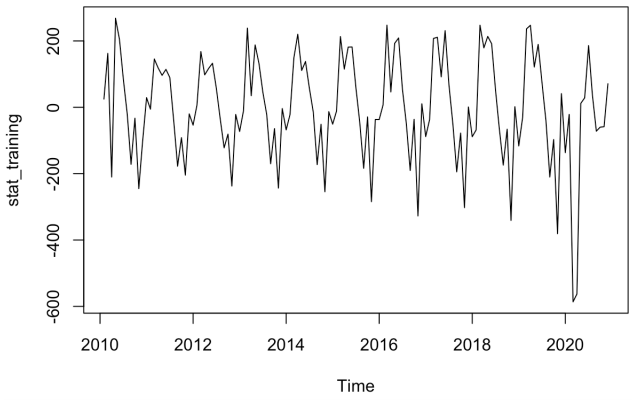

Fig. 16: The stat_training Time Series

The ARIMA models are combinations of AR(p) models, in which AR stands for autoregressive, and MA(q) models, in which MD stands for moving average . If the original time series is necessary to have a differencing procedure, the I(d) model, in which I stands for integrated, needs to be considered to the combination.

AR(p) models are used to predict current value based on the past value in a time series, whose formula is defined as below:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{y-p} + \varepsilon_t$$

In which p is called the order of the AR model which indicates the number of past or lagged values are used to predict the current value.

While MA(q) models are different from the AR(p) models, which are using the past forecast errors to predict the current value:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

In which q indicates the number of lagged errors or called shocks which effect the predication of current value.

When the differencing is combined with AR and MA model, a non-seasonal ARIMA model is generated. This full model:

$$\begin{aligned} y'_t = c &+ \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \ldots + \phi_p y'_{y-p} \\ &+ \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \end{aligned}$$

In which, y' is a new time series which is generated by applying differencing to the original series.

Hence, ARIMA models are used to report the autocorrelations of a time series, in which "autocorrelation" means the correlations of the corresponding series with it own past values. And SARIMA models are seasonal ARIMA.

In this case the differencing frequency has been determined as d = 1. The difficult part is determine the value p and q in AR(p) and MA(q) models. There are two functions are used to determine the value p and q. And they are functions to represent the autocorrelation, they are the autocorrelation function (ACF) and the partial autocorrelation function (PACF). However, the results plot of these two functions some times give more model options, see Fig. 17 and 18. In order to determine the value of p and q and identify a ideal ARIMA/SARIMA model to fit the ts_training time series, a auto.arima(ts_training) function[1] is used to derive the model directly, see Fig. 19.

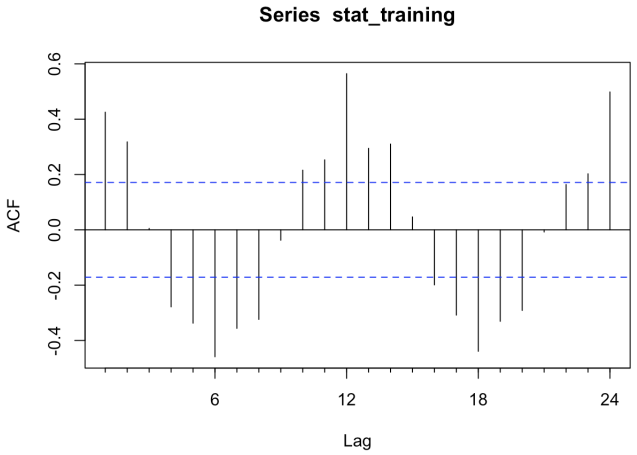

Fig. 17: A correlogram of stat_training

Two approaches are used to test weather the selected model is optimum. One is to check the residuals or error terms are normally and independently distributed, see Fig. 20. The other is using Box-Ljung test, the result, p-value is not significant, indicates that the autocorrelations don't differ from zero and the selected model is fit the training dataset well (Fig. 21).

The accuracy of this model is shown in Fig 22, in particular, the RMSE value is 100.11 which is less than the corresponding RMSE value of the ETS model. Hence the SARIMA model is considered as the best model to predict the departures values in the future. Using such model to forecast the departures of the first 6 months in 2021, see Fig. 23 and the plot of prediction is in Fig. 24.

C. Simple Time Series Models

Several simplest forecasting methods are listed below with their name and the corresponding function respectively.

[1]If applying this function to the stat_training series, the model is selected as ARIMA$(1,0,0)(0,1,1)[12]$, which indicates the differencing of the original time series has been done. Other coefficients are exactly equvalent.

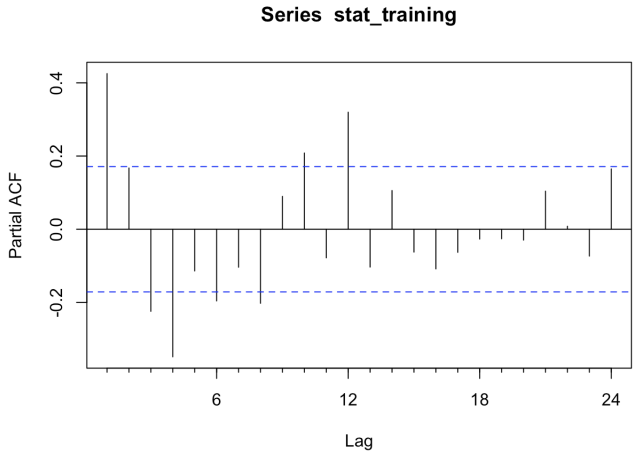

Fig. 18: A Partial Autocorrelation of stat_training

```
> fit_auto_ts_training <- auto.arima(ts_training)
> fit_auto_ts_training
Series: ts_training
ARIMA(1,1,0)(0,1,1)[12]

Coefficients:
          ar1     sma1
      0.3966  -0.8274
s.e.  0.0837   0.2045

sigma^2 = 11307:  log likelihood = -730.08
AIC=1466.16   AICc=1466.37   BIC=1474.5
```

Fig. 19: An automated ARIMA model of ts_training

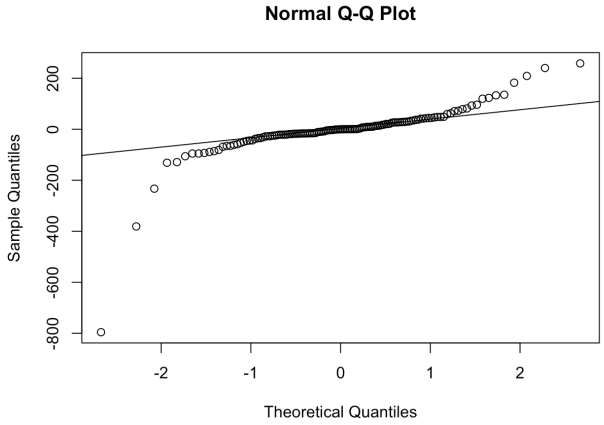

Fig. 20: The Linear Fit of Residuals

```
> Box.test(fit_auto_ts_training$residuals, type = "Ljung-Box")

        Box-Ljung test

data:  fit_auto_ts_training$residuals
X-squared = 0.00067123, df = 1, p-value = 0.9793
```

Fig. 21: The Box-Ljung Test of the SARIMA Model

```
> round(accuracy(fit_auto_ts_training),2)
                     ME   RMSE   MAE    MPE  MAPE MASE ACF1
Training set -4.24 100.11 50.25 -18.05 35.48 0.26    0
```

Fig. 22: The Accuracy of the SARIMA Model

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2021	104.45103	-32.29759	241.1996	-104.6879	313.5900
Feb 2021	94.42007	-140.31920	329.1593	-264.5827	453.4228
Mar 2021	168.66266	-147.82503	485.1504	-315.3635	652.6888
Apr 2021	181.99916	-203.89686	567.8952	-408.1779	772.1762
May 2021	312.76668	-133.51826	759.0516	-369.7672	995.3006
Jun 2021	466.67683	-33.35034	966.7040	-298.0487	1231.4024

Fig. 23: The Forecast of The First 6 months in 2021 using SARIMA Model

- Average method, all future values are the same to the mean of the past values, which is accomplished by using function $\mathsf{meanf}(y, h)$.
- Naïve method (Random Walk), the prediction values identical to the last observation, which is implemented by using function $\mathsf{naive}(y, h)$.
- Seasonal naïve method, the forecasting value equals to the last observation of the same season, which is achieved by using function $\mathsf{snaive}(y, h)$.
- Drift method using function, is built based on the naïve method. The average change between the prediction values and the historical values in the naïve method are generated, and these changes called the drift. And it can be executed by using the function $\mathsf{rwf}(y, h, \mathsf{drift} = \mathsf{TRUE})$. By applying these simple methods to the original departure time series, the result plot (Fig. 25) present the prediction by using different simple forecasting models.

Finally, a linear trend model has been created by setting the Departures as the dependent variable and the Months as the independent variable, see Fig 26. By using the Durbin-Watson Test, this linear trend model does not suitable to forecast the corresponding sample dataset, since the result is significant and residuals or error terms are autocorrelated.

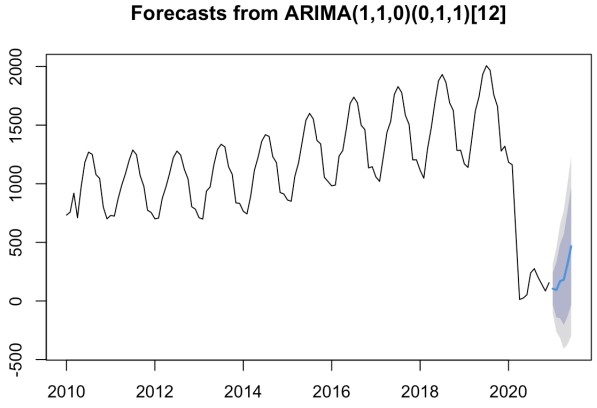

Fig. 24: The Plot of Forecast using SARIMA Model

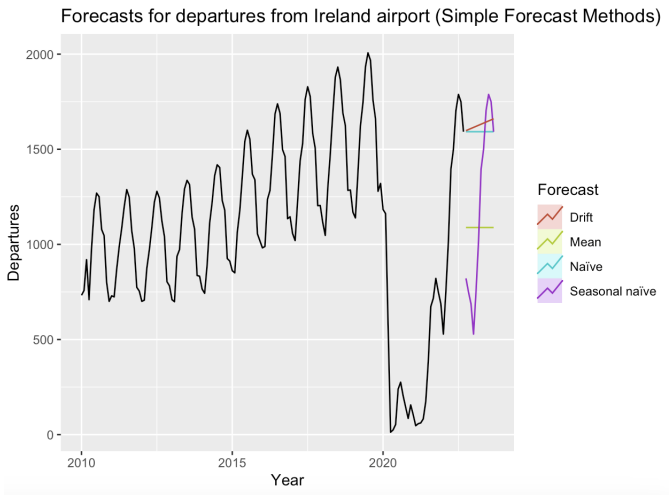

Fig. 25: The Simplest Forecast of Departures

```
> lt_model <- lm(Departures~Months, data = df_departures)
> summary(lt_model)

Call:
lm(formula = Departures ~ Months, data = df_departures)

Residuals:
     Min       1Q   Median       3Q      Max
-1060.19  -309.85    44.95   298.08   930.91

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1273.46043  480.03000   2.653  0.00883 **
Months        -0.01092    0.02828  -0.386  0.69983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 470.2 on 151 degrees of freedom
Multiple R-squared:  0.0009872, Adjusted R-squared:  -0.005629
F-statistic: 0.1492 on 1 and 151 DF,  p-value: 0.6998

> durbinWatsonTest(lt_model)
 lag Autocorrelation D-W Statistic p-value
   1       0.9356174      0.116029       0
 Alternative hypothesis: rho != 0
```

Fig. 26: The Durbin-Watson Test on a Linear Trend Model

V. Logistic Regression Model

A. Dataset Description and Data Pre-processing

A dataset, which contains 45211 entries and 17 variables, has been provided in CSV format, whose name is bank.csv. The values in bank.csv presents a bank's marketing campaign details of some different clients. The demand of the bank is to persuade customers to purchase a bank product, by creating a logistic regression model based on this dataset. By validation of the data, there is no NULL value, no duplicated observations in this dataset. Data has been extract and read from the CSV file. The first 6 rows this dataset are list to show the structure of data, see Fig. 27, in which y is the dependent variable and other 16 variables are considered as predictors initially. Any variable has only "yes" or "no" values, has transformed to the corresponding variable with values 1 and 0 respectively.

In order to understand the meaning of variables more clearly, categorical values of some variables are displayed in Fig. 28.

```
> head(df_bank)
  age          job marital education default balance housing loan
1  58   management married  tertiary       0    2143       1    0
2  44   technician  single secondary       0      29       1    0
3  33 entrepreneur married secondary       0       2       1    1
4  47  blue-collar married   unknown       0    1506       1    0
5  33      unknown  single   unknown       0       1       0    0
6  35   management married  tertiary       0     231       1    0
  contact day month duration campaign pdays previous poutcome y
1 unknown   5   May      261        1    -1        0  unknown 0
2 unknown   5   May      151        1    -1        0  unknown 0
3 unknown   5   May       76        1    -1        0  unknown 0
4 unknown   5   May       92        1    -1        0  unknown 0
5 unknown   5   May      198        1    -1        0  unknown 0
6 unknown   5   May      139        1    -1        0  unknown 0
```

Fig. 27: The Bank Data Details

```
> unique(df_bank$job)
 [1] "management"    "technician"    "entrepreneur"  "blue-collar"
 [5] "unknown"       "retired"       "admin."        "services"
 [9] "self-employed" "unemployed"    "housemaid"     "student"
> unique(df_bank$marital)
[1] "married"  "single"   "divorced"
> unique(df_bank$education)
[1] "tertiary"  "secondary" "unknown"   "primary"
> unique(df_bank$contact)
[1] "unknown"   "cellular"  "telephone"
> unique(df_bank$month)
 [1] "may" "jun" "jul" "aug" "oct" "nov" "dec" "jan" "feb" "mar"
[11] "apr" "sep"
> unique(df_bank$poutcome)
[1] "unknown" "failure" "other"   "success"
```

Fig. 28: Categorical Values of Some Variables

B. Principle Components Analysis

There are 16 independent variables in this dataset, however, some of them has strong correlations from each other. By identifying such variables and removing them from the predictors group, the complexity of the statistical model will be decreased and the model is more understandable. A dimension-reduction technique, named Principal components analysis(PAC), is used to realise this requirement. In this case, all numeric variables are considered, from the Scree plot, Fig. 29, of these variables, the number of principle components or factors is 3, in other words, 9 variables are linear combined into 3 components, see Fig. 30. And these three components will be new predictors to build a logistic regression models rather than 9 variables. Variables *pday* and *previous* are related, both are used to present the previous contact information before this campaign, hence a new component which combines them together is named *pre_campaign*. Similarly, variables *compaign*, *day* and *duration* are all express the contact information during this campaign, a new components which group them together is called *dur_campaign*. The last new component, which is a linear combination of four variables *balance*, *loan*, *default* and *housing*, is given a name *finance*.

So far, a new dataset (see Fig. 31) has been created by adding 3 new components *finance*, *pre_compaign* and *dur_compaign*, while deleting 9 variables which were used

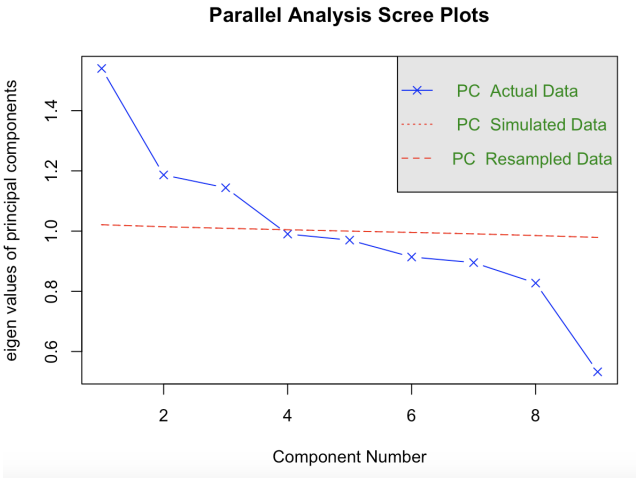

Fig. 29: The Eigenvalues for the Principal Components.

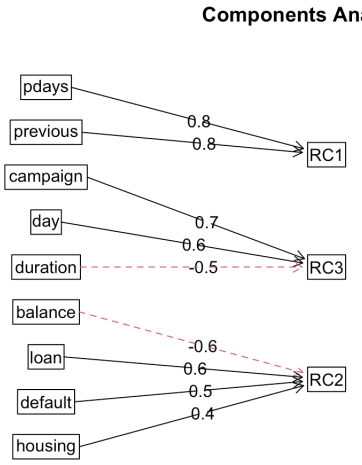

Fig. 30: The Grouping of Principle Components

to combined these 3. This new dataset is ready for building a logistic regression model.

```
> head(df_bank_pca)
  age          job marital education contact month
1  58   management married  tertiary unknown   may
2  44   technician  single secondary unknown   may
3  33 entrepreneur married secondary unknown   may
4  47  blue-collar married   unknown unknown   may
5  33      unknown  single   unknown unknown   may
6  35   management married  tertiary unknown   may
  poutcome    finance pre_compaign dur_compaign y
1  unknown -0.9796261   0.23088787   -0.5714997 0
2  unknown -1.0104449   0.19903206   -0.1333177 0
3  unknown -1.5686439   0.07400683    1.1895146 0
4  unknown -0.8119431   0.19645555   -0.3093768 0
5  unknown -0.7564545  -0.13786833   -0.7521365 0
6  unknown -0.9781765   0.19796114   -0.1536321 0
```

Fig. 31: A New Dataset after PCA.

C. Building Logistic Regression Model Step by Step

Before creating the logistic regression model, dataset has been split into a training set contains 36169 observations and

a test set which is composed of 9042 entries. Dichotomous dependent variable y is decisions of clients on this bank product in compaign, hence a binary logistic regression model, called bankfit, has been built based on the training dataset (see Fig. 32). However, before examining the coefficients of

```
bankfit<-glm(y~age+job+marital+education+contact+month+poutcome
             +finance+pre_compaign+dur_compaign
             ,data =training, family=binomial)
```

Fig. 32: A Logistic Regression Model based on Training Dataset

each variable in this model, a statistical test for individual predictors, called variable importance test (see Fig. 33), has been applied to this model to demonstrate the significance of each predictor to this model. The bigger numbers show the corresponding predictor is more import to the response. The variable with smaller importance value is considered to be cut off. Hence, 4 alternative models have been created by getting rid of these variables, variables *age*, *job*, *marital* and *education*, one by one. In order to identify the most appropriate model, a Pseudo R^2 [10] test and an ANOVA test have been used to compare the alternative models one by one with the original model.

```
> #Variable Importance
> varImp(bankfit)
                      Overall
age                 0.6144076
jobblue-collar      3.6175607
jobentrepreneur     1.4919744
jobhousemaid        3.1057573
jobmanagement       0.7291889
jobretired          3.4433113
jobself-employed    0.8173273
jobservices         1.5220993
jobstudent          3.6978312
jobtechnician       1.3187876
jobunemployed       0.2941081
jobunknown          0.3514046
maritalmarried      3.0603103
maritalsingle       1.8047162
educationsecondary  1.9297016
educationtertiary   3.8191565
educationunknown    0.9849532
contacttelephone    1.7230079
contactunknown     16.6424519
monthaug           10.1639088
monthdec            1.8049148
monthfeb           11.0990571
monthjan            3.4079523
monthjul            7.5104564
monthjun            2.9510171
monthmar            8.8144985
monthmay           10.3879759
monthnov           10.1080702
monthoct            7.1446602
monthsep            2.6117483
poutcomeother       4.4631822
poutcomesuccess    26.2143471
poutcomeunknown     7.8736300
finance            31.6092992
pre_compaign        1.2940319
dur_compaign       34.8123671
```

Fig. 33: Variable Importance

By applying the Pseudo R^2 test to the bankfit model and a model without *age* variable, named bankfit1 model. The result (see FIg. 34) shows the change does not have big effect on the original model. The ANOVA test, by comparing the likelihood ratio of these two models, (see Fig. 35) presents the similar result that cutting the *age* variable is not significant to the original model. Other optional models, for example, bankfit2 model cut off two predictors, varNameage and *marital*, has a vital impact to the original model, are not selected.

The logistic Regression Model Equation:

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

```
> pR2(bankfit)
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden
-9.989007e+03 -1.307826e+04  6.178500e+03 2.362127e-01
          r2ML          r2CU
 1.570293e-01  3.050352e-01
> pR2(bankfit1)#age
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden
-9.989196e+03 -1.307826e+04  6.178123e+03 2.361983e-01
          r2ML          r2CU
 1.570205e-01  3.050182e-01
```

Fig. 34: Comparison between Models using Pseudo R^2 Test.

```
> anova(bankfit, bankfit1, test ="Chisq")
Analysis of Deviance Table

Model 1: y ~ age + job + marital + education + contact + month + poutcome +
    finance + pre_compaign + dur_compaign
Model 2: y ~ job + marital + education + contact + month + poutcome +
    finance + pre_compaign + dur_compaign
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     36132      19978
2     36133      19978 -1  -0.3772   0.5391
```

Fig. 35: Comparison between Models using ANOVA.

In which, the coefficients β_i and the corresponding predictors x_i $(i = 0, 1, \ldots p)$ are listed in the Fig. 36. For instance, $\beta_0 = -1.16524$, $\beta\_1 = -0.27283$ and x_1 is *jobblue-collar* and so on. And there are 35 independent variables, $p = 35$.

D. Evaluation and Cross-Validation

By applying the logistic regression model bankfit1 to test dataset, a cross-validation has been completed by creating a confusion matrix, see Fig. 37. The accuracy indicates that the model is fit the dataset very well.

VI. Conclusion and Future Work

Time series analysis and logistic regression model are critical statistical prediction approaches. Models, which have been created in this paper, are fit very well on their dataset respectively. For time series analysis, the SARIMA model has smaller RMSE, has been considered as the model. However, the ETS model also presented a good performance. Both model forecasted that the number of departures from Ireland airport will increase. For the logistic regression model has been created based on the bank details, shows a positive fitting to the dataset. The future work is to create these models by using Machine Learning algorithms. And models and test in this paper all implemented in R language, hence, using SPSS software application to build these models and analysis their accuracy is considered.

Acknowledgment

I would like to express our grateful thanks to my lecturers Hicham Rifai and Tony Delaney for their patience and detailed explanation. I learned statistics 20 years ago, without their guidance, I cannot reach here. I would like to thank my husband and my son for their supports.

```
> summary(bankfit1)

Call:
glm(formula = y ~ job + marital + education + contact + month +
    poutcome + finance + pre_compaign + dur_compaign, family = binomial,
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0772  -0.4573  -0.3137  -0.1856   4.5721

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.16524    0.13624  -8.553  < 2e-16 ***
jobblue-collar    -0.27283    0.07498  -3.639 0.000274 ***
jobentrepreneur   -0.18560    0.12621  -1.471 0.141413
jobhousemaid      -0.43779    0.14262  -3.070 0.002143 **
jobmanagement     -0.05355    0.07591  -0.705 0.480565
jobretired         0.37405    0.08998   4.157 3.23e-05 ***
jobself-employed  -0.09230    0.11499  -0.803 0.422134
jobservices       -0.13111    0.08536  -1.536 0.124536
jobstudent         0.40902    0.11216   3.647 0.000266 ***
jobtechnician     -0.09380    0.07145  -1.313 0.189257
jobunemployed      0.03517    0.11414   0.308 0.757988
jobunknown        -0.07655    0.23941  -0.320 0.749171
maritalmarried    -0.18815    0.06031  -3.120 0.001810 **
maritalsingle      0.11014    0.06490   1.697 0.089679 .
educationsecondary 0.12270    0.06538   1.877 0.060543 .
educationtertiary  0.28724    0.07615   3.772 0.000162 ***
educationunknown   0.10970    0.11023   0.995 0.319645
contacttelephone  -0.12533    0.07599  -1.649 0.099072 .
contactunknown    -1.17595    0.07070 -16.633  < 2e-16 ***
monthaug          -0.78722    0.07759 -10.146  < 2e-16 ***
monthdec           0.33817    0.18698   1.809 0.070514 .
monthfeb          -0.95984    0.08656 -11.089  < 2e-16 ***
monthjan          -0.41790    0.12279  -3.403 0.000665 ***
monthjul          -0.57505    0.07654  -7.513 5.78e-14 ***
monthjun          -0.26963    0.09144  -2.949 0.003192 **
monthmar           1.14329    0.12951   8.828  < 2e-16 ***
monthmay          -0.75903    0.07299 -10.399  < 2e-16 ***
monthnov          -0.86607    0.08579 -10.095  < 2e-16 ***
monthoct           0.81927    0.11412   7.179 7.03e-13 ***
monthsep           0.33326    0.12682   2.628 0.008594 **
poutcomeother      0.41334    0.09273   4.457 8.30e-06 ***
poutcomesuccess    2.23356    0.08517  26.225  < 2e-16 ***
poutcomeunknown   -0.77505    0.09825  -7.888 3.06e-15 ***
finance           -0.80958    0.02561 -31.607  < 2e-16 ***
pre_compaign      -0.05388    0.04132  -1.304 0.192259
dur_compaign      -0.76541    0.02188 -34.979  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26157  on 36168  degrees of freedom
Residual deviance: 19978  on 36133  degrees of freedom
AIC: 20050

Number of Fisher Scoring iterations: 6
```

Fig. 36: The Summary of the Selected Logistic Regression Model.

```
> pred_test<-predict(bankfit1, test, type="response")
> pred_class_test<-ifelse(pred_test<0.5, "0","1")
> bought_test<-as.factor(test$y)
> pred_classf_test<-as.factor(pred_class_test)
> confusionMatrix(bought_test,pred_classf_test,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7861  136
         1  805  240

               Accuracy : 0.8959
                 95% CI : (0.8895, 0.9022)
    No Information Rate : 0.9584
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2947

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.63830
            Specificity : 0.90711
         Pos Pred Value : 0.22967
         Neg Pred Value : 0.98299
             Prevalence : 0.04158
         Detection Rate : 0.02654
   Detection Prevalence : 0.11557
      Balanced Accuracy : 0.77270

       'Positive' Class : 1
```

Fig. 37: Cross Validation of the Model using Confusion Matrix

REFERENCES

[1] N. Nurhamidah, N, Nusyirwan and A. Faisol, “Forecasting seasonal time series data using the holt-winters exponential smoothing method of additive models.” Jurnal Matematika Integratif, 16(2), pp.151-157, 2020.

[2] Z. Li, J. Bi, and Z. Li, “Passenger flow forecasting research for airport terminal based on SARIMA time series model.” In IOP conference series: earth and environmental science (Vol. 100, No. 1, p. 012146). IOP Publishing, December 2017.

[3] H. Lee, W. Malik and Y. C. Jung, “Taxi-out time prediction for departures at charlotte airport using machine learning techniques.” In 16th AIAA Aviation Technology, Integration, and Operations Conference (p. 3910), 2016.

[4] H. Zhou, W. Li, Z. Jiang, F. Cai and Y. Xue, “Flight Departure Time Prediction Based on Deep Learning.” Aerospace, 9(7), p.394, 2022.

[5] K. Annin, M. Omane-Adjepong, and S. S. Senya, “Applying logistic regression to e-banking usage in Kumasi Metropolis, Ghana. ” International Journal of Marketing Studies, 6(2), p.153, 2014.

[6] T. Zaghdoudi “Bank failure prediction with logistic regression”. International Journal of Economics and Financial Issues, 3(2), pp.537-543, 2013.

[7] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing.” Decision Support Systems, 62, pp.22-31, 2014.

[8] E. S. Gardner, and E. McKenzie, “Why the damped trend works.” Journal of the Operational Research Society, 62(6), pp.1177-1180, 2011.

[9] R.J. Hyndman and G. Athanasopoulos “Forecasting: principles and practice”, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. 7.3 Holt-Winters’ seasonal method, 2018 [Online Available: https://otexts.com/fpp2/holt-winters.html] Accessed on 050123

[10] G.A.J. Hemmert, L.M. Schons, J. Wieseke, and H, Schimmelpfennig “Log-likelihood-based Pseudo-R2 in Logistic Regression: Deriving Sample-sensitive Benchmarks” Sociological Methods & Research Vol. 47(3) 507-531, 2018.