Time Series Analysis on Dublin Airport Departure Dataset & Logistic Regression on Marketing campaign Dataset - TABA

Sureshkumar Durairaj
ID: 21178933
MSc in Data Analytics - B – 2022-2023
Statistics for Data Analytics
Terminal Assignment-Based Assessment – Semester 1
National College of Ireland, IRELAND
Email: x21178933@student.ncirl.ie

***Abstract*: The increased innovation in airline and aircraft technologies has increased the usage of air traffic over the years, a wide variety of data has been accumulated across various departments. Hence, we have taken a dataset with number of departures from Ireland airport through the period 2010 to 2022 using time series analysis, specifically using Simple Time series methods, Exponential smoothing & ARIMA / SARIMA Models. The digital campaign has been a key facilitator for the upliftment of the business, banking industry thrives and continues to improve with its continuous indulgence with the customers physically and digitally. In view of the same we are considering a digital campaign data from a bank which has the attributes aiming to acquire successful sale of the bank product indicated by a binary classifier. Hence, Logistic regression a simple rather an effective modelling technique to predict the accuracy and effectiveness of the prediction.**

Keywords — Time Series Analysis, ARIMA / SARIMA, Exponential smoothing, Accuracy, Logistic Regression

I. INTRODUCTION

The inception of digital technology has given rise to the surplus flow of data from all businesses and domains. Therefore, we need to employ an effective mechanism for analyzing the DNA of the data and understanding the pattern or trends to discover KPI groups thereby helpful in organizing the data for effective usage and strategizing to cater the business decisions. Statistics is a broader study of data patterns which helps the data analysts to understand the detailed insights of data.

The data in general are of 3 different types of time series (partly regressive), regression data and classification data (binomial or polynomial). These data are driven by the target variables (Y) and the independent variables (X). For effective analysis of each of these types of data we can employ a broad variety state-of-the-art data mining and machine learning techniques. For our analysis we are considering 2 different data sets

II. DATA SET DESRIPTION

A. Departure count – Dublin from 2010 to 2022
This data set contains the time series of number of departures from Ireland via Irish airports, which consists of the data from year 2010 to 2022 month wise.

B. Banking – Marketing Campaign Data[8]
This data set consists of the marketing campaigning data from a Bank which has the details of the key attributes and the outcome of the campaign determining the acceptance of a customer into buying a banking product.

III. RELATED WORKS

We begin our research by going through the existing work on the similar datasets. In this paper regarding the time series analysis [1], they have considered the airport traffic data from commercial ADS-B data providers for analysis. The airport traffic could be influenced by numerous movements such as airplanes, skydivers, helicopters, bird flock and so on. These are rather easier to obtain owing to slow movement and hence accumulated and clustered together for analysis. These are then statistically processed to get aggregate values of various properties such as sum of ground times by each time dimension , count of airport connections for each aircraft and the departure/arrival time of these aircrafts in all the airports by time and so on These are then brought in to a time series and generated visualizations such as heat maps to plot the contours of proximities of the individual aircraft and cluster map of the particular day/hour which effects in aggressive traffic scenarios . Here they have employed neural network to arrive at a

refined model to generate the prediction of collision in an attempt to minimize the same.

In the paper [2] , They have considered Xiamen Gaoqi International Airport in China for the analysis , this consist of the data and a Fundamental diagram which lead to the discovery of key features such as relative velocity ,trajectory similarity and flight displacement . Here in they are employing a KNN algorithm using the same they are classifying the air traffic into 4 types of Free Flow, Transitional Flow, slightly congested flow and heavily congested flow. The time series which are arrived from these data are then transformed into complex network through the GUI method. They are then analyzed in terms of degree of distribution and structure of the network established. in case of FF those nodes with higher degree evaluates into a higher cluster and the ones with lesser is solitary. The nodes in the TF however have almost similar degrees distributed across the network. Those with SCF is smaller than the TF and HCF even smaller. Although this method is effective in terms of arriving at the dynamics of air traffic this can further be improved by build network subsets which can be supported by extraction of further insights from the air traffic flow.

In the paper [3] regarding the time series analysis, they are considering the data set which has the prediction of Checked in Baggage departed from Airport terminal using time series analysis. Here they are employing SARIMA plot since the idea is to arrive at an effective technique to help facilitate the airport management in resource allocation depending on the predicted count. They are considering the data from Kunming Changshui International airport for the analysis. An empirical analysis is carried to estimate the best model by considering only important numeric attributes and time bound seasonal data to arrive at an effective model. The mean value error of the finalized model is between 23 to 26. From the results it can be described that the usage of SARIMA has resulted in the long-term prediction. The RSME value is between 28 to 35 and the relative RSME is between 2.27 to 2.5respectively. However, this can further be improved by including the heavy loaded seasonal data such as holiday season data and there is no clear indication of the terminal which can provide further insights.

In the paper [4] related to the marketing campaign, They are considering the telemarketing data from banking sector . The customer response is accumulated and provisioned as a delimited file which can be used for analysis using hybrid machine learning models. The following are the models used for the analysis such as KNN, Random Forest, SVM, Naïve Bayes, Logistic Regression and extremely randomized trees. They employ EDA to improve the prediction of higher subscription rate. Here they use an ensemble classifier in order to address the imbalance in the data. The accuracy achieved using the Random Forest is about 94.02%, the same using Logistic Regression is 79.34%, accuracy percent arrived using Decision tree is 92.80%, the same using SVM is 82.79%, prediction accuracy obtained using the Naïve Bayes is 73.66% and KNN is 87.00%. From these we can conclude that the Random Forest has the highest accuracy in comparison to the other models used.

In the paper [5] on the marketing campaign data, the same data in use for our analysis is being used and they have employed a SMOTE approach for arriving a model in order to estimate the prediction performance. Here they also consider the minority classes unlike the other usual analysis which focuses only on the majority classes. A total of 150 features are analyzed and finalized data set is used for the analysis. Here they use Naïve Bayes for the analysis which gives 100% accuracy on “0” class data and 0% accuracy on class “1” which is not reliable. Hence, we can consider to use the other effective modelling technique such as Logistic regression

IV. PRE-PROCESSING

A. Departure count – Dublin from 2010 to 2022

The given data consist of 2 columns and 153 rows one for each month from 2010 to 2022 as shown in the Fig 1 below,

```
dat <- read_csv("Departure.csv")
glimpse(dat)

Rows: 153 Columns: 2
── Column specification ────────
Delimiter: ","
chr (1): Month
dbl (1): departures '000

ℹ Use `spec()` to retrieve the full c
ℹ Specify the column types or set

Rows: 153
Columns: 2
$ Month            <chr> "2010 Jar
$ `departures '000` <dbl> 732.4,
```

Fig 1 – Data set Details

For the sake of clarity, we have renamed the 2nd column as” departure” and converted the 1st column into date format by concatenating it with ‘01’ therefore getting a date in the

YYYY-MM-DD format. Then we remove the old field and convert the data into a time series for our analysis; The data frame after converting into time series as shown in the Fig 2 below,

A Time Series: 13 × 12

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	De
2010	732.4	757.2	919.6	709.5	977.9	1183.1	1269.5	1250.5	1078.7	1045.8	800.7	700.
2011	729.6	724.0	869.6	987.9	1084.2	1198.5	1288.0	1247.8	1070.3	978.7	774.3	754.
2012	700.5	706.9	874.8	972.6	1089.6	1222.3	1278.8	1244.2	1122.1	1041.3	804.1	782.
2013	709.5	698.7	937.3	972.3	1160.2	1292.1	1336.7	1314.3	1144.2	1079.9	836.5	832.
2014	764.7	742.8	892.5	1112.7	1223.8	1361.8	1418.4	1403.9	1231.2	1179.9	925.4	912.
2015	861.5	850.3	1063.4	1178.5	1359.9	1541.7	1600.0	1552.5	1368.8	1339.8	1055.4	1018.
2016	981.7	989.0	1236.5	1282.8	1475.7	1684.6	1738.7	1688.9	1498.6	1462.3	1134.9	1145.
2017	1057.1	1020.2	1227.7	1438.7	1530.6	1761.5	1828.6	1778.1	1583.6	1505.9	1203.5	1204.
2018	1115.6	1047.3	1294.6	1473.8	1687.4	1879.0	1932.0	1865.2	1691.0	1624.9	1284.0	1285.
2019	1169.3	1139.1	1375.4	1622.6	1744.0	1933.4	2006.9	1967.5	1757.4	1660.1	1279.1	1320.
2020	1183.5	1161.9	575.6	12.8	24.7	53.1	239.0	275.5	203.5	143.7	85.1	156.
2021	104.1	46.7	57.5	61.5	82.4	174.7	384.0	673.0	716.6	821.2	749.7	687.
2022	528.4	751.9	1014.7	1395.1	1499.1	1704.0	1788.2	1749.6	1592.2	732.4	757.2	919.

Fig 2 Time Series of the given Data set

Then to begin with our data cleaning on the data set we need to ensure there are no abnormalities or null values in the give data set. The given Fig 3 shows that the data set has no null values,

```
sum(is.na(dat))
```
```
0
```

Fig 3 Null value check

B. Banking – Marketing Campaign Data[8]

Secondly, we are considering the data set which contains the information about the marketing campaign which has the list of features which evaluates to result in the target variable(y) containing 2(binomial) values. The given data set has about 45211 records and 17 attributes.

```
bank_dat <- read_csv("bank.csv")
```
```
Rows: 45211 Columns: 17
── Column specification ──

Delimiter: ","
chr (10): job, marital, education, default, housing, loan, contact, month, p...
dbl  (7): age, balance, day, duration, campaign, pdays, previous

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this messag
```

Fig 4 Bank – Marketing Campaign data overview

As with any data set, here also we begin by pre-processing the given data. First, we validate the data to identify whether there are any NULL values. From the Fig 4 below we can see that there are no NULL or NA values in the given data set,

```
any(is.null(bank_dat))
```
```
FALSE
```
```
any(is.na(bank_dat))
```
```
FALSE
```

Fig 4 NULL Data check on the time series

The data wise detail on the given data set is showcased using a bar graph as shown in the below Fig 5

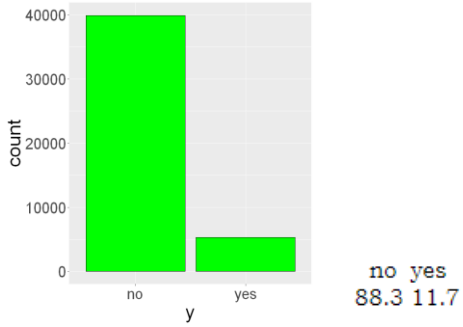


Fig 5 Target Variable composition

V. EVALUATION & ANALYSIS

A. Departure count – Dublin from 2010 to 2022

Let us begin the analysis of the time series data by using a simple plot () which shows the linearity of the departure count in Ireland airports as shown in Fig 6 below. As per the initial observation the seasonality of the data is pretty evident.

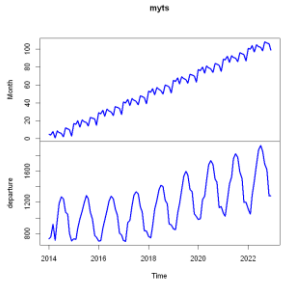


Fig 6 Passenger count vs Year

Hence the seasonal plot of the time series on the given data set. From the plot shown in the Fig 7, the departure counts are exponential with consistent constant values in each period in the time series plotted month wise to observe the pattern in a different time dimension other than year.

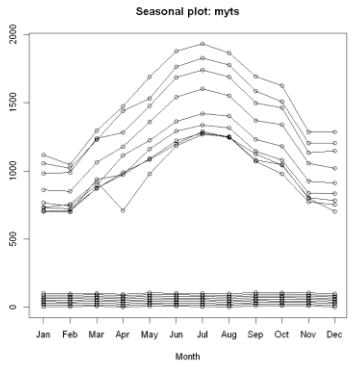


Fig 7 Seasonal Plot of the time series

Applying model on the given time series data. We apply a linear model on the given data shows which shows the linearity trend of the data over the years from the plot as shown in the Fig 8 below,

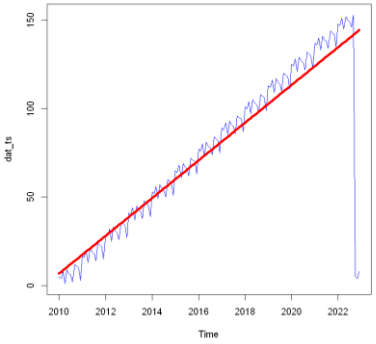


Fig 8 Linear Model Plot

The scatter plot of the model generated is as shown in the Fig 9 below,

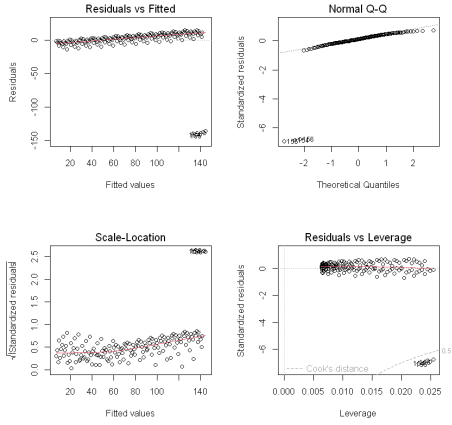


Fig 9 Scatter Plot of Linear Model

Furthermore, we need to analyze the data by factorizing the data between -2 to 2 to enforce the consistency between the monthly data as depicted in Fig 8 Left. Also, the STL decomposition using Loess is shown on the right in Fig 8 below.

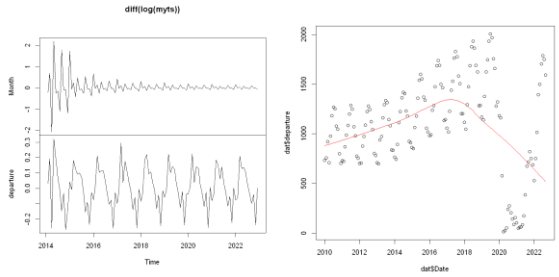


Fig 8 Plot on differential log

We now apply the linear model on the log of the time series to determine the pattern as shown in the Fig 9 below,

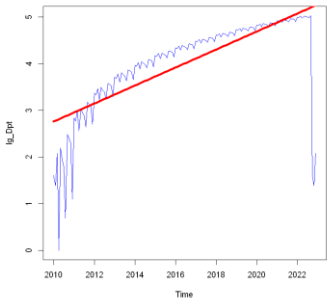


Fig 9 Linear model on Log (time series)

The boxplot per month shows us the pattern of the yearly cycle without reference to non-standard libraries.

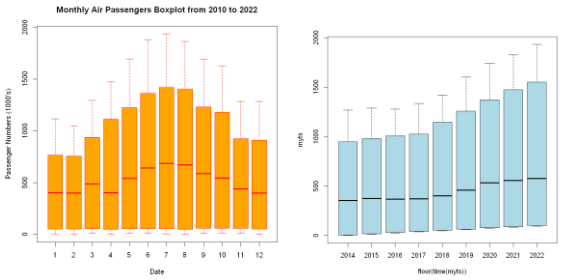


Fig 10 Box plot on the given data set

The decomposition of the additive time series can be performed using plot on the decompose () on the time series data as shown in Fig 11 below, which explains a consistent pattern on the seasonal plot and linear

trend and sudden dip and rise at the fague end of the data set.

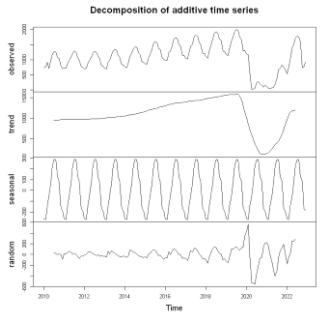
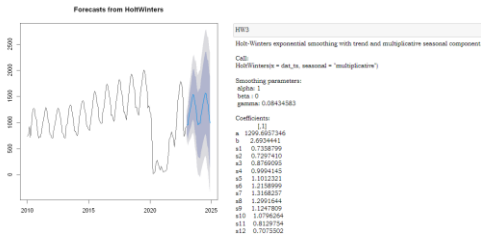


Fig 12 Decomposition of data

Then we use different modelling on the time series obtained to identify the best method for the same. The forecast form the Holt winters method is as shown in the Fig 13(a) below, also on the right side in the Fig 13(b) below illustrates the summary of Holt-winters method. The elaborated plot on the multiplicative seasonal component is captured in the Fig 14 below,



(a) *(b)*

Fig 13 Forecast from HW finalized

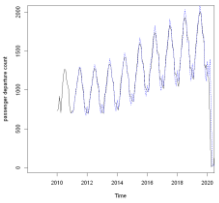


Fig 14 Holt winters Multiplicative Seasonal Component & Trend

Then we perform the hypothesis testing on the arrived forecast using the holt winter's method using acf() a shown in the Fig 15 below , this is not a performant method as it is clearly evident that some of the Lag values are outside the boundary .

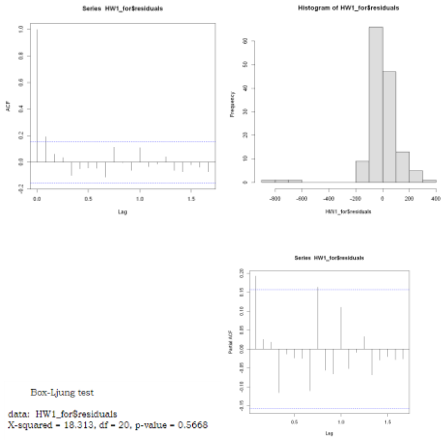


Fig 15 Hypothesis Testing on the Time Series

For comparison we can also perform an alternative hypothesis on the actual time series as shown in the Fig 16 below,

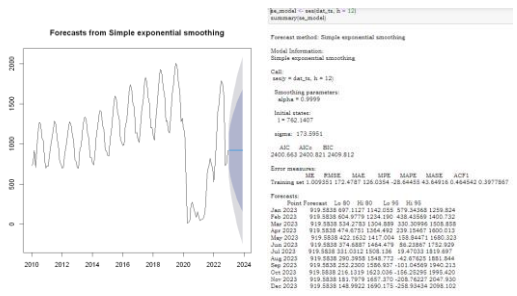
```
adf.test(dat_ts)

        Augmented Dickey-Fuller Test

data:  dat_ts
Dickey-Fuller = -2.3689, Lag order = 5, p-value = 0.4226
alternative hypothesis: stationary
```

Fig 16 Alternative Hypothesis Testing

The forecast using SET is shown in the Fig 17 below, which highlights the projection for over 2022 in yellow. The RSME value (172) is slightly higher as shown below.



The MAPE value arrived from the Simple exponential model is 170.369663484254

Fig 17 Forecast from SET

The model arrived using ARIMA seems to have the better performance as it has reduced the outfitters as shown in the Fig 18 below

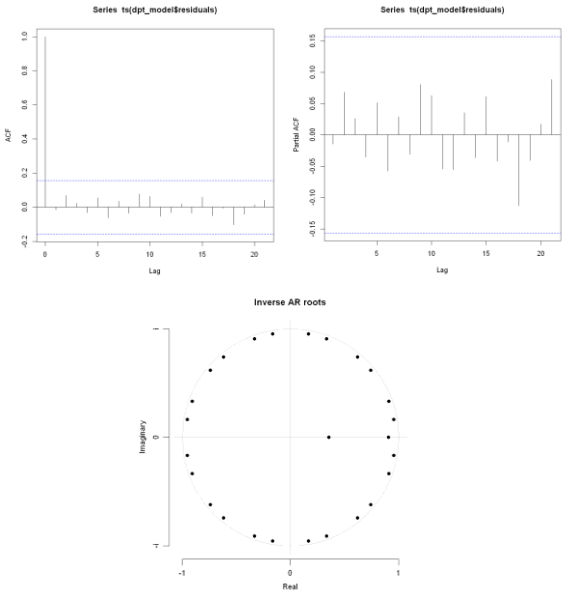
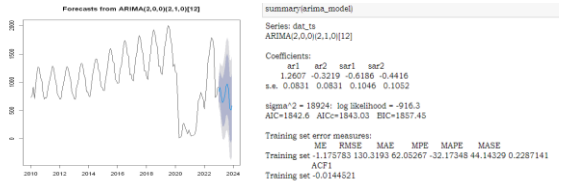


Fig 17 Improvements from ARIMA model using acf() and pacf()

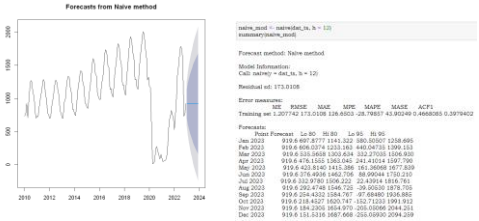
As shown in the Fig 18 below, the RSME value arrive d using the ARIMA model is the least compared to th e other models as shown in the upcoming discussion.



The MAPE value from the forecast of ARIMA model - 143.157754891103

Fig 18 Forecast from ARIMA

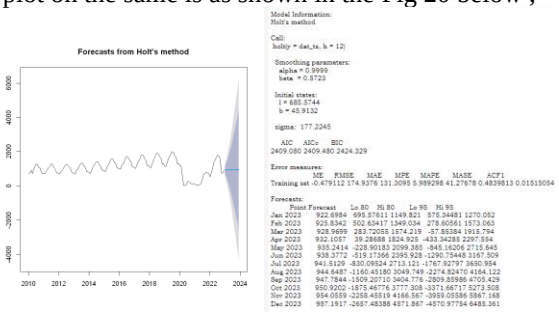
To perform a comparative analysis using different mo dels we consider using Naïve Method and the forecast results are displayed in the Fig 19,



MAPE value calculated from the forecast of Naïve model is 170.37203881612

Fig 19 Naïve Method – Model forecast

As seen above the RSME value is increasing 173.010 8 , so the performance is not on the rise . Hence we c ould anlyze other models . Let us now consider to us e the Holt method , unlike the holt winters method w hich considers seasonality and the trend in the predict ions here in the holt method we could perform a linea r exponential smoothing . The resulting model and th e plot on the same is as shown in the Fig 20 below ,



MAPE value for the Holt model is 172.492479221163

Fig 20 Forecast from Holt method

From the above we could observe that the RSME valu e is 174 and obviously the performance has not steppe d up. Hence, we can conclude that the ARIMA model yielded the highest performance.

Forecast from Jan 2021 to June 2021 using the decisiv e model which in our case is ARIMA. From the given data set the time series used for the ARIMA model is then filtered from 2010 to 2020 as shown in the Fig 2 1 below, which is then subjected to forecast the next 6 months using the intervals property.

A Time Series: 11 × 12

	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	1299.5	1250.5	1078.7	1045.8	800.7	700.5
2011	729.6	724.0	899.6	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	851.5	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	1885.2	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

Fig 20 Time series of the Filtered data

Then we apply the Arima model and plot the same to determine the outcomes which is showcased in the Fi g 21 below,

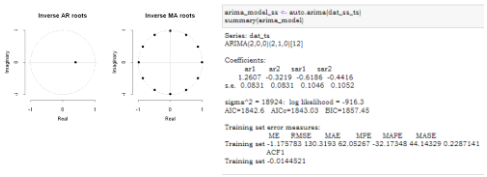


Fig 21 ARIMA Model and Forecast on the filtered time series

The MAPE value obtained using the restrained time s eries model is 104.6 and the RSME value is 130.3193 which is pretty good from the initial observations. Al so, for comparison we can analyze the same using the second best model which is the SES, the forecasting f or the same is as shown in Fig 22 below,

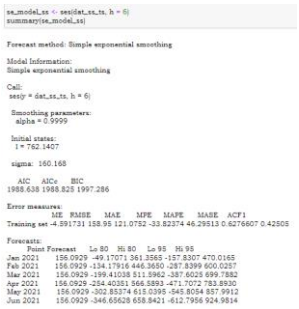


Fig 22 SES Forecasting on Filtered time series

From the above we could infer that the performance f rom SES is inferior comparing to the ARIMA. So we could now compare the forecasted data arrived from ARIMA with the source data set, which is shown in t he Fig 23 below.

fore_arima_2020

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2021	104.45102	-32.29760	241.1996	-104.6879	313.5900
Feb 2021	94.42005	-140.31922	329.1593	-264.5827	453.4228
Mar 2021	168.66260	-147.82510	485.1503	-315.3636	652.6888
Apr 2021	181.99907	-203.89695	567.8951	-408.1780	772.1761
May 2021	312.76658	-133.51837	759.0515	-369.7673	995.3005
Jun 2021	466.67672	-33.35045	966.7039	-298.0488	1231.4023

```
2021 January,104.1
2021 February,46.7
2021 March,57.5
2021 April,61.5
2021 May,82.4
2021 June,174.7
```

Fig 23 Comparison of Forecast vs Source Data

Although not perfect the forecast using the ARIMA m odel seems to be optimal in comparison to the other m odels for the give data set as observed from the above comparison, where initially the predictions were to th e point which began to dip as the trend goes on furthe r.

B. **Banking – Marketing Campaign Data**[8]
The snapshot of the data set for the analysis is as shown in the Fig 24 below,

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no
32	blue-collar	single	primary	no	23	yes	yes	unknown	5	may	160	1	-1	0	unknown	no
25	services	married	secondary	no	50	yes	no	unknown	5	may	342	1	-1	0	unknown	no
40	retired	married	primary	no	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no

Fig 24 Bank – Marketing Campaign Data set

The attributes for building a model can be analyzed using the correlation plot which will indicate the key columns that will contribute to the effectiveness of the model, this is illustrated in the Fig 25 below,

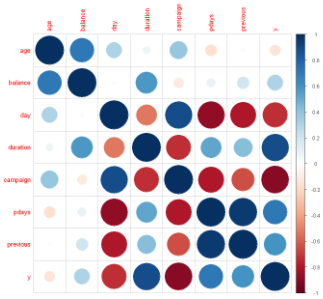


Fig25 Correlation Plot on the given dataset

For simplicity let us initiate the analysis by understanding the composition of the different attributes in the given data set. The plot in the Fig 26(a) shows the count of users age wise who has taken this campaign.

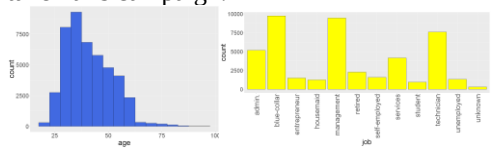


Fig 24 Count of users vs Age(a) & Occupation(b)

Similarly, the count of users by occupation is as shown in the Fig26(b) above. The count comparison of the users by Other attributes (Marital_status,Education & defaulter) is as shown in the below Fig 27 (a), (b) &(c) respectively.

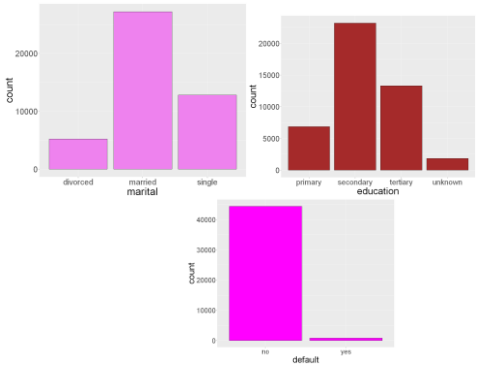


Fig 27 Count of users by (Marital_status,Education & defa ulter)

Since the Money is key factor in determining the mar keting analytics it is important to derive the pattern on the bank balance of the individuals participating in th e campaign to establish the accuracy of prediction wh ich is depicted in the Fig 28 below,

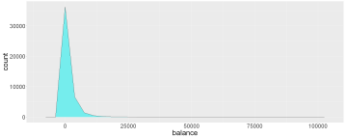


Fig 28 Count of users vs their avg. Balance

This gives us a clear idea that these are potential custo mer who could be affected into buying a bank product .

In the same way we have analyzed the other attributes as shown in the Fig 29 below,

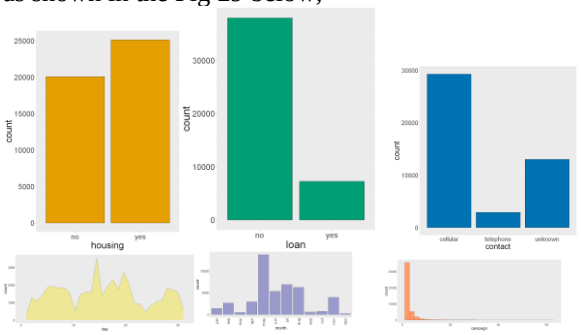


Fig 29 Count of users vs Key features

Finally, we consider rather an important field for effective d ecision making which is the previous out come to determine the effectiveness of the campaign which is as shown below Fig 30 .

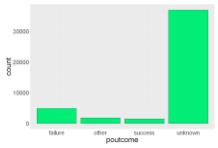


Fig 30 Previous outcome distribution by count

Instead of beginning the analysis by sampling them, it is a good way to begin the model building with the ac tual raw data itself the summary of the same is shown in the Fig 31 below ,

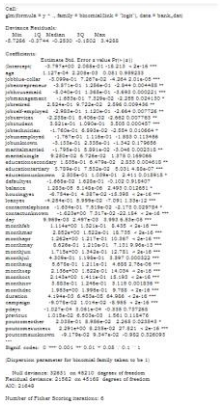


Fig 31 Raw Data – Logistic Regression – Summary

The effectiveness of the Logistic regression model ca n be interpreted by the value of the Area Under ROC curve, which is as shown in the Fig 32 below,

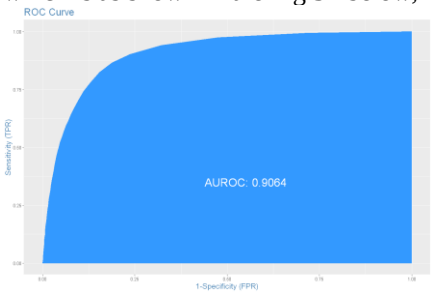


Fig 32 ROC Curve for Raw data Log Reg Model

To measure the accuracy of the prediction in case of classifi cation model the best method is to use a confusion matrix o n the actual vs predicted which is as depicted in the Fig 33 b elow,

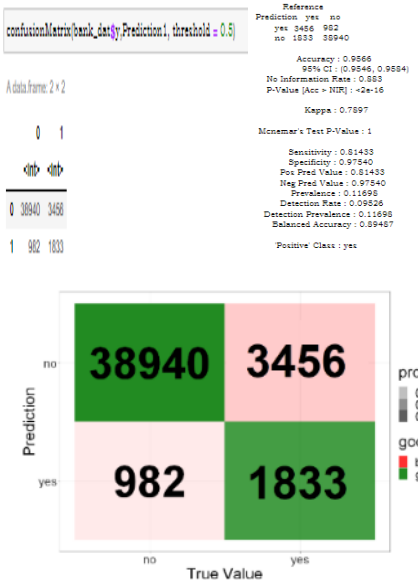


Fig 33 Confusion Matrix – Raw Data Log Reg and its sum mary

The accuracy achieved is about 95 % but it is not eno ugh to conclude the effectiveness of the prediction, he nce we continue to enhance the model by including sa mpling, 80:20 approach is considered here to build th e dataset for analysis as shown in the Fig 34 below,

```
set.seed(8)
train.size = floor(0.8*nrow(log_rg))
train.index = sample(1:nrow(log_rg), train.size)
train.set = log_rg[train.index,]
test.set = log_rg[-train.index,]
```

Fig 34 Sampling and Split – Initial

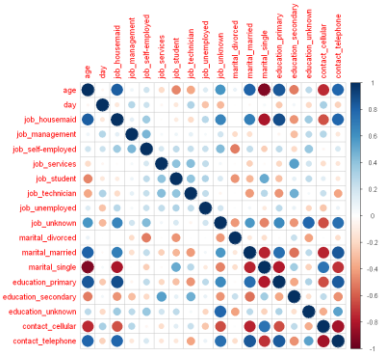


Fig 35 Correlation plot on the enhanced data set

The above plot shows the dependency of each of the i ndependent variable against the dependent variable on the enhanced data set as shown in the Fig 36. Here the Log_rg is the data set which is derived by excluding t he unimportant independent variables which will cont ribute to the effectiveness of the prediction.

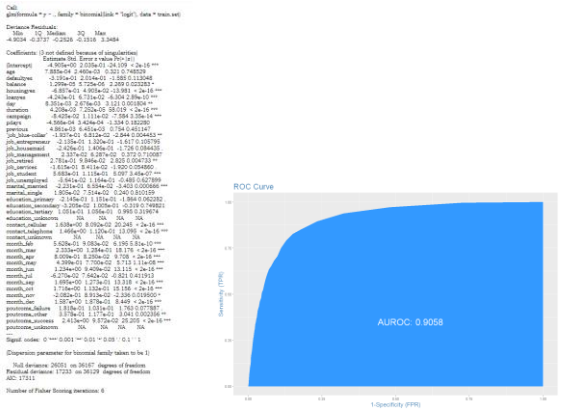


Fig 36 Log Reg using 1st sampling and ROC curve

From the above Fig 36 we can see the summary of the Log. Reg. Model generate and the ROC curve has a sl ightly reduced value which indicates the improvemen t in performance in comparison to the previous model . Similarly, we measure the prediction accuracy by co nfusion matrix as shown in the Fig 37 below ,

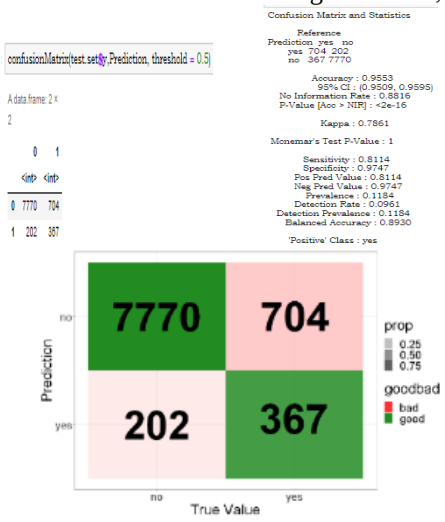


Fig 37 Confusion Matrix – 2nd Log Reg Model and its sum mary

From the above although it is looking like the same accurac y of 95% there is slight increase about 0.003 in the enhance model arrived .

```
optCutOff = optimalCutoff(test.set$y, Prediction, optimiseFor = "misclasserror")
optCutOff
```

```
0.409999922514258
```

Fig 38 Optimal Cut off Threshold

An optimal cut off value in the above Fig 38 is arrive d to accurately plot the ROC curve which yielded the same value of AUROC as 0.9058 as the original plot u sing the current model. Finally, we attempt to further i mprove the model by considering only the numeric co lumns in the data set by filtering other vector columns and sampled the same using a slightly modified 75:25 approach. The summary of the Logistic regression mo del is as illustrated in the Fig 39 below,

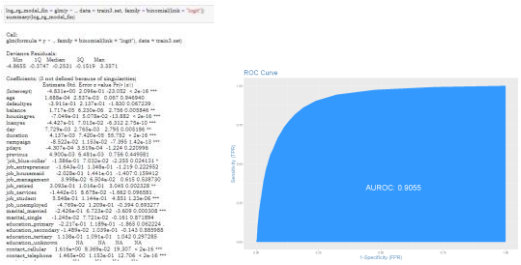


Fig 39 Final Log. Reg. Model & its ROC Curve

The value of AUROC has been reduced to 0.9055 wh ich is the better of all the models arrived so far . Henc e the accuracy of the same can measure by the confus ion matirx as shown in the Fig 40 below,

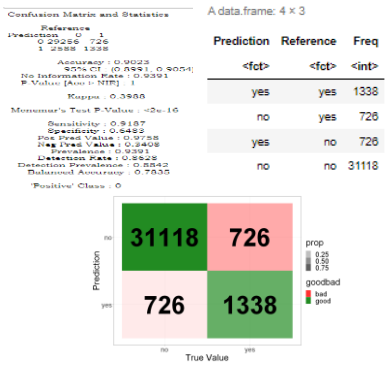


Fig 40 Confusion Matric of final Log. Reg model & its sum mary

This shows that the accuracy of the finalized model (a fter performing the dimension reduction) using the lo gistic regression is close to 96% which proves to be th e efficient of all the other models generated. Hence, w e can ignore the previous models and consider this as the finalized model.

B) (I) CHECK FOR ASSUMPTI ONS – LOGISTIC REGRESSION

1) From the given dataset the target variable y (which had 2 values(dichotomous)yes / no w hich are then factorized to 0 or 1 hence it ca n be concluded that the finalized Log. Reg Model is mutually exclusive.
2) Sample Size: The data set is having a large v alue about 45211 records and the sampling d ata since it is considered about 75% of the a ctual (33908 records), this data set is a perfe ct fit
3) Non-Multicollinearity: From the correlation plot in Fig 36, we can confirm that the attrib utes considered for the model generation is multi collinear since it depicts a clear relatio nship between the 11 to 15 variables, and so me does not have any relationship.

VI. CONCLUSION & SUMMARY

From our analysis on the airport departure count in Ireland by Month data after performing the analysis on the various time series model such as Simple Exponential Smoothing, Naïve Bayes, Holt winter, Holt method and ARIMA we could say that the ARIMA model yielded the highest performance with least RMSE value of 130, hence the same is used to predict the forecast for data for first 6 months of 2021. Similarly, using the logistic regression we have analyzed the banking marketing campaign and after improving the model by considering different combinations of correlated independent variables the accuracy achieved is around 96% .However , these two data sets can further be analyzed in detail using Neural networks for time series to yield a heatmap to provide efficient insights , obtain less RSME value and use other techniques such as Random forest , KNN and SVM by effective sampling and wisely choosing the independent variables . The code used for the analysis is available in the link in [6] and [7] respectively.

VII. REFERENCES

1) Dästner, K., Schmid, E., zu Roseneckh-Köhler, B. V. H., & Opitz, F. (2020, October). Analysis of Time Series of Statistical Air Traffic Data. In *2020 21st International Radar Symposium (IRS)* (pp. 157-162). IEEE.
2) Li, S., Wang, C., & Wang, J. (2020). Exploring dynamic characteristics of multi-state air traffic flow: A time series approach. *IEEE Access*, *8*, 64565-64577.
3) Saeed, S. E., Hammad, M., & Alqaddoumi, A. (2022, March). Predicting Customer's Subscription Response to Bank Telemarketing Campaign Based on Machine learning Algorithms. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1474-1478). IEEE.
4) Ma, Q., Bi, J., Sai, Q., & Li, Z. (2021, July). Research on Prediction of Checked baggage Departed from Airport Terminal Based on Time Series Analysis. In *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)* (pp. 264-269). IEEE.
5) Islam, M. S., Arifuzzaman, M., & Islam, M. S. (2019, December). SMOTE Approach for Predicting the Success of Bank Telemarketing. In *2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)* (pp. 1-5). IEEE.
6) https://studentncirl-my.sharepoint.com/:u:/r/personal/x21178933_student_ncirl_ie/Documents/Stats_TABA.ipynb?csf=1&web=1&e=3jqQlv
7) https://studentncirl-my.sharepoint.com/:u:/r/personal/x21178933_student_ncirl_ie/Documents/Bank_LogisticReg.ipynb?csf=1&web=1&e=WhCbGy
8) S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014