# 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](mailto: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 a irport 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

1. 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

1. DATA SET DESRIPTION
2. 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.

1. 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.

1. 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

1. PRE-PROCESSING
2. 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,

Text

Description automatically generated

*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,

Table

Description automatically generated

*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,

Graphical user interface

Description automatically generated with medium confidence

*Fig 3 Null value check*

1. 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.

Graphical user interface, text, application, email

Description automatically generated

*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,

Graphical user interface, text, application, email

Description automatically generated

*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

Chart, bar chart, histogram

Description automatically generated Text

Description automatically generated

Fig 5 Target Variable composition

1. EVALUATION & ANALYSIS
2. **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.

Chart, histogram

Description automatically generated

*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.

*Diagram

Description automatically generated*

*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,

*Chart

Description automatically generated*

*Fig 8 Linear Model Plot*

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

*Diagram

Description automatically generated*

*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.

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated

*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,

*Chart

Description automatically generated*

*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.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

*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.

Chart

Description automatically generated

*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,

*Chart, scatter chart

Description automatically generated*Text, email

Description automatically generated

1. *(b)*

*Fig 13 Forecast from HW finalized*

*Chart

Description automatically generated*

*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 .

*Chart

Description automatically generatedChart, histogram

Description automatically generated*

Text, letter

Description automatically generated *Chart, box and whisker chart

Description automatically generated*

*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,

Graphical user interface, text, application, email

Description automatically generated

*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.

Chart, scatter chart

Description automatically generatedText

Description automatically generated

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

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart

Description automatically generated

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

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

*Chart, scatter chart

Description automatically generated* Text

Description automatically generated

The MAPE value from the forecast of ARIMA model - 143.157754891103

*Fig 18 Forecast from ARIMA*

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

Chart, scatter chart

Description automatically generated Text

Description automatically generated

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 173.0108 , so the performance is not on the rise . Hence we could anlyze other models . Let us now consider to use the Holt method , unlike the holt winters method which considers seasonality and the trend in the predictions here in the holt method we could perform a linear exponential smoothing . The resulting model and the plot on the same is as shown in the Fig 20 below ,

Chart, histogram

Description automatically generatedText

Description automatically generated

MAPE value for the Holt model is 172.492479221163

*Fig 20 Forecast from Holt method*

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

Forecast from Jan 2021 to June 2021 using the decisive 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 21 below, which is then subjected to forecast the next 6 months using the intervals property.

Table

Description automatically generated

*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 Fig 21 below,

Chart, engineering drawing

Description automatically generatedText

Description automatically generated

*Fig 21 ARIMA Model and Forecast on the filtered time series*

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

Text

Description automatically generated

*Fig 22 SES Forecasting on Filtered time series*

From the above we could infer that the performance from 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 the Fig 23 below.

Text

Description automatically generated

Text

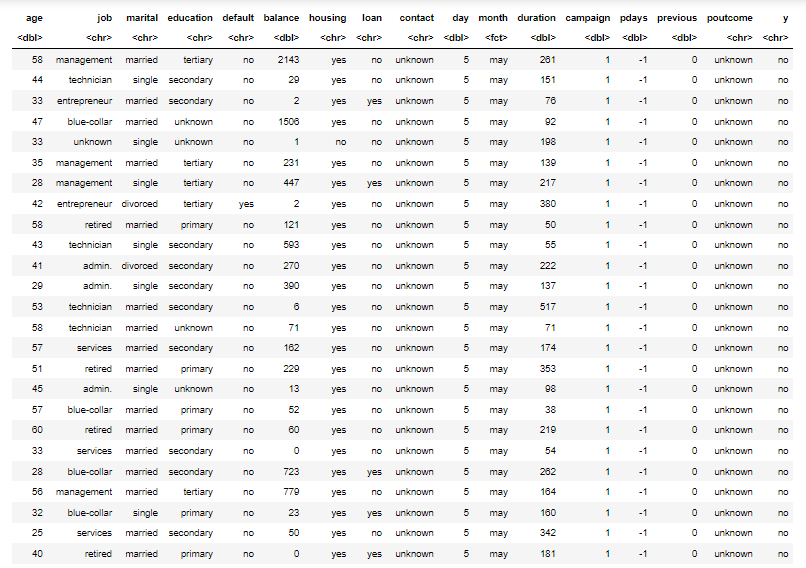
Description automatically generated

*Fig 23 Comparison of Forecast vs Source Data*

Although not perfect the forecast using the ARIMA model seems to be optimal in comparison to the other models for the give data set as observed from the above comparison, where initially the predictions were to the point which began to dip as the trend goes on further.

1. **Banking – Marketing Campaign Data**[8]

The snapshot of the data set for the analysis is as shown in the Fig 24 below,



*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,

Chart, bubble chart

Description automatically generated

*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.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

*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.

*Chart, bar chart

Description automatically generated*Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

*Fig 27 Count of users by (Marital\_status,Education & defaulter)*

Since the Money is key factor in determining the marketing analytics it is important to derive the pattern on the bank balance of the individuals participating in the campaign to establish the accuracy of prediction which is depicted in the Fig 28 below,

Chart

Description automatically generated

*Fig 28 Count of users vs their avg. Balance*

This gives us a clear idea that these are potential customer 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,

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated Chart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generated Chart

Description automatically generated

*Fig 29 Count of users vs Key features*

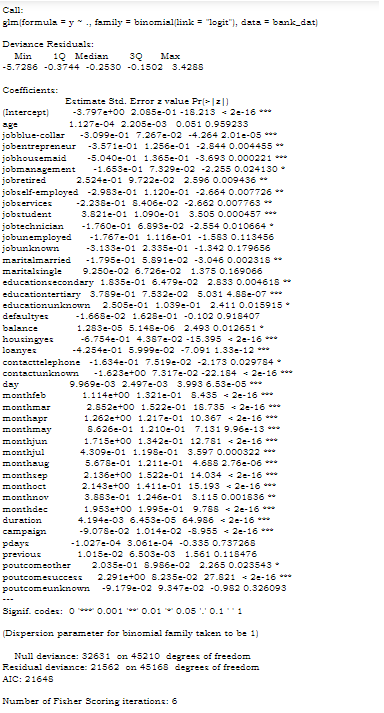
Finally, we consider rather an important field for effective decision making which is the previous out come to determine the effectiveness of the campaign which is as shown below Fig 30 .

Chart, histogram

Description automatically generated

*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 actual 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 can be interpreted by the value of the Area Under ROC curve, which is as shown in the Fig 32 below,

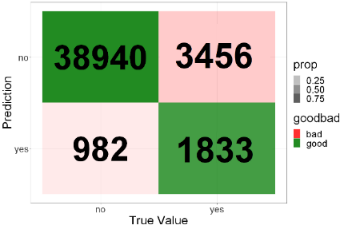
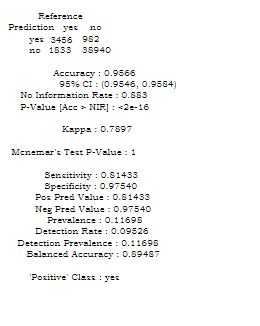
Chart, treemap chart

Description automatically generated

*Fig 32 ROC Curve for Raw data Log Reg Model*

To measure the accuracy of the prediction in case of classification model the best method is to use a confusion matrix on the actual vs predicted which is as depicted in the Fig 33 below,

Graphical user interface, text, application

Description automatically generated

*Fig 33 Confusion Matrix – Raw Data Log Reg and its summary*

The accuracy achieved is about 95 % but it is not enough to conclude the effectiveness of the prediction, hence we continue to enhance the model by including sampling, 80:20 approach is considered here to build the dataset for analysis as shown in the Fig 34 below,

Text

Description automatically generated

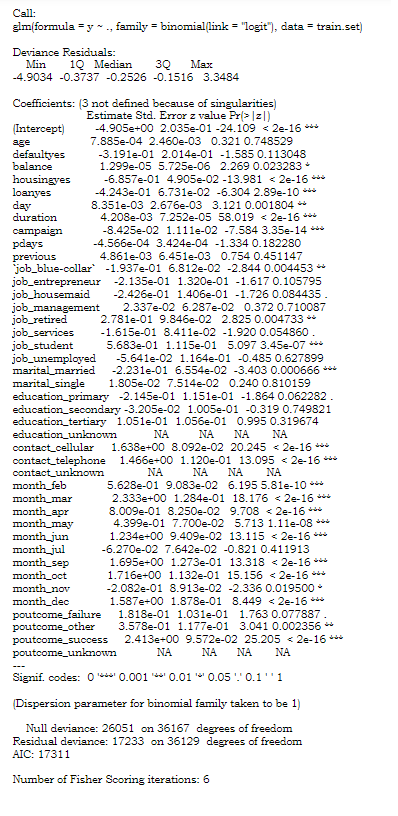
*Fig 34 Sampling and Split – Initial*

*Chart, scatter chart

Description automatically generated*

*Fig 35 Correlation plot on the enhanced data set*

The above plot shows the dependency of each of the independent 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 the unimportant independent variables which will contribute to the effectiveness of the prediction.

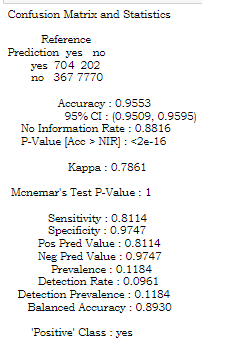
 Chart, treemap chart

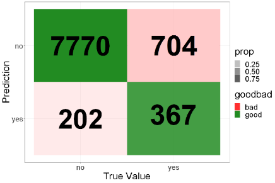
Description automatically generated

*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 slightly reduced value which indicates the improvement in performance in comparison to the previous model. Similarly, we measure the prediction accuracy by confusion matrix as shown in the Fig 37 below ,

Graphical user interface, text, application, email

Description automatically generated 



*Fig 37 Confusion Matrix – 2nd Log Reg Model and its summary*

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

Graphical user interface, text, application

Description automatically generated

*Fig 38 Optimal Cut off Threshold*

An optimal cut off value in the above Fig 38 is arrived to accurately plot the ROC curve which yielded the same value of AUROC as 0.9058 as the original plot using the current model. Finally, we attempt to further improve the model by considering only the numeric columns 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 model is as illustrated in the Fig 39 below,

Text

Description automatically generated Chart, treemap chart

Description automatically generated

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

The value of AUROC has been reduced to 0.9055 which is the better of all the models arrived so far . Hence the accuracy of the same can measure by the confusion matirx as shown in the Fig 40 below,

Text, letter

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceA picture containing logo

Description automatically generated

*Fig 40 Confusion Matric of final Log. Reg model & its summary*

This shows that the accuracy of the finalized model (after performing the dimension reduction) using the logistic regression is close to 96% which proves to be the efficient of all the other models generated. Hence, we can ignore the previous models and consider this as the finalized model.

1. (I) CHECK FOR ASSUMPTIONS – LOGISTIC REGRESSION
2. From the given dataset the target variable y (which had 2 values(dichotomous)yes / no which are then factorized to 0 or 1 hence it can be concluded that the finalized Log. Reg Model is mutually exclusive.
3. Sample Size: The data set is having a large value about 45211 records and the sampling data since it is considered about 75% of the actual (33908 records), this data set is a perfect fit
4. Non-Multicollinearity: From the correlation plot in Fig 36, we can confirm that the attributes considered for the model generation is multi collinear since it depicts a clear relationship between the 11 to 15 variables, and some does not have any relationship.
5. 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.

1. REFERENCES
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. <https://studentncirl-my.sharepoint.com/:u:/r/personal/x21178933_student_ncirl_ie/Documents/Stats_TABA.ipynb?csf=1&web=1&e=3jqQlv>
8. <https://studentncirl-my.sharepoint.com/:u:/r/personal/x21178933_student_ncirl_ie/Documents/Bank_LogisticReg.ipynb?csf=1&web=1&e=WhCbGy>
9. 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