Study on Data Analysis of Postal Section Routine and its improvement using Machine Learning

Rajasekaran.T.A[1*], Vijayalakshmi. P[2*], V.Rajendran[3]
[1] Research Scholar, [2]Asst. Professor, [3]Professor & Director
[1,2,3,]Department of Electronics And Communication Engineering, Vels institute of Science, Technology and Advance Studies (VISTAS), Chennai.
*tarajasekaran@gmail.in
*viji.se@velsuniv.ac.in

Abstract: This Indian Post Office acts as a traditional saving mode for many section of society especially to rural households. It is reported that the postal sections have it's routine movement of business, That data collected on daily basis will be useful to improve its business. The project aims at collecting data on regular postal section routine like sale of stamps, speed post booked and delivered, money-order, postal saving deposits, parcel movement,PLI/RPLI etc.From the data it is proposed to carry out supervised ML process to find its behavioral business clusters. The main goal of this thesis is to develop a machine learning model that could predict impact of COVID-19 using suitable algorithm and to develop such a model for the various products of department of post, a literature study alongside an experiment is set to identify a suitable algorithm to assess the features that impact the prediction model.

Key Words: COVID-19, Machine Learning, Prediction, Supervised Learning, Classification Techniques.

I. INTRODUCTION

For more than 150 years Department of Post served as a backbone for communication throughout the nation, correspondingly developed the habit of savings and investment among people through their Post Office Savings Banking Schemes and thereby aiding the socio-economic development of the country. Even the savings and investments are practiced from early 19th century; it has been become the vital and common one in the present scenario where it focuses for the purpose of future security. Post Office Savings and Investment Schemes are like the commercial banking schemes, which offer several opportunities for the investors and mobilize the savings of people with relatively small income with massive platform in handling future hand techniques in serving its customers. In current situation the impact of corona virus disease 2019 (COVID-19) has caused pandemic situation in every aspect of life and organization working pattern.

This creates a challenge to use machine learning techniques for the generating the current impact caused by the disease in account opening format and the statistical anomalies are not considered. Also, the appropriate selection of parameters and the selection of the best machine learning model for prediction are other challenges involved in training a model.

In this project, we are going to perform Supervised learning using two best traditional algorithm for time series prediction they are Random forest and ARIMA to predict next 12 months data with the changes caused by COVID-19 first wave in every small saving scheme

product of the Postal Organization, the study impact on some parameters such as economic statistics, lockdown patterns from the data collected from the site https://mis.cept.gov.in/

This site has collection of data posted on daily routine from Pan India ,dealing with various sectors of the postal section, small saving schemes data was posted from September 2020 onwards, the influencing factors in the performance of financial models studies were made for the last financially year starting from September 2020 ending with March 2021,The result concludes the scope of the impact created by the expected versus the predicted dataset, using best time series algorithm suing the comparison for developing accurate predictive model, thus creating seeds for promising future studies.

II. RELATED WORK

Account opening prediction is an essential task which has to be done by Department of Posts and the prediction will be able to provide crucial impact to towards the business decision making process. Not only that by having income prediction for the e-commerce platform, they can have a better understanding about their financial status to manage the workforce, and further improving their supply chain management system. The prediction of account opening allows understanding the lifecycle of the saving schemes platform as its product and growth, stability, decline and how are the openings during the COVID19 period

Prediction is usually done by using the most common method, time series analysis. Time series analysis involve the Autoregressive function which helps which any type of prediction analysis. According to [1] study on the machine learning model for sales times series forecasting, it has mention that sales prediction is a modern business intelligence method. Also mention by [3], ARIMA model has a better approach for the performance in prediction in the time series analysis. This main problem stated in this research by [1] is that for time series data, the data required is large to capture the seasonality and the large transactional sales data can have many missing data and outliers. These data will then need to be take into account a lot of different factors which can impact the sales. The goal for this time series analysis it to combine different time series algorithm in order to improve this prediction accuracy.

There were five algorithm which have been selected in the research which are Extra Tree, ARIMA, Random Forest, Lasso and Neural Network which are all time series algorithm and supervised. Based on the results for the forecasting error testing, Extra Tree has the highest validation error compared to the rest and Neural Network has the lowest validation error making it one of the best algorithms for prediction. Based on the other researcher [3], have conducted a research on forecasting of Walmart sales using machine-learning algorithms. The key for this research was done by implementing several different classification algorithms in the sales data from all different Walmart locations all over the United States. The problem which was highlighted in this research was creating a competitive comparative analysis to find the best algorithm. The researcher had selected 3 different algorithms for the comparison and test it using the MAE. The goal of this research is to find accuracy of algorithm using different hyper parameters of each model to obtain the best Mean Absolute

Error (MAE) and forecasting error (RMSE). The algorithm which was used for this research was Random Forest.

The results of this research [3] indicate that the Random Forest is the best algorithm, which have scored the minimum amount in MAE evaluation (1979.4) which have shown a high accuracy compared with the others. Another research was conducted by [2], which was to study the sales forecast for Amazon sales based on the different statistic methodology. This research has primarily focused on the Amazon data and forecast the future sales using the historical data by using statistic algorithms. The goal for this research is to conduct sensitivity analysis on the two methods, and identify which is the most reliable, accurate and suitable approach. The better the accuracy for a method, the better will the prediction will be for the sales forecasting. There were three different approaches used for this research which includes Winters' exponential smoothing, time-series decomposition and ARIMA. The results of this research were done by measuring the forecasting error (RMSE).

All of the method has a very low amount of forecasting error, therefore all of the method can be implemented to conduct sales forecasting for the Amazon sales. Another research which was done by [5], which discuss about Explaining Machine Learning models in sales prediction. This research mainly discusses about all the main models of machine leaning which is commonly being used in sales prediction and also will show analysis of the best machine learning model available. The problem which is identified by this research paper was that how to identify the appropriate model based on the business understanding by using the intelligence and data driven models. The goal of this research was to demonstrate how effective each of the model is and its usability. This is being done to ensure that the correct method have been selected to the selected business environment was mentioned by [6] and [7]. The method (algorithm) which was been used by this research was the decision tree, neural network, naïve bayes, random forest and support vector machine (SVM).

Based on the chosen algorithm, the results are tabulating against the accuracy. The random forest is at 85%, naïve bayes is at 83%, decision tree at 76%, neural network at 70% and finally the SVM is at the lowest 59%. Therefore, the best method to be chosen is the random forest as it has a high accuracy model at 85%. Another research done by [8] where Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks all over the world , ARIMA and Random Forest time series models to incidence data of outbreaks of highly pathogenic avian influenza (H5N1) in Egypt, available through the online EMPRES-I system. We found that the Random Forest model outperformed the ARIMA model in predictive ability the Random Forest model is effective for predicting outbreaks of H5N1 in Egypt. Conclusions: Random Forest time series modeling provides enhanced predictive ability over existing time series models for the prediction of infectious disease outbreaks. This result, along with those showing the concordance between bird and human outbreaks (Rabinowitz et al. 2012), provides a new approach to predicting these dangerous outbreaks in bird populations based on existing, freely available data. Our analysis uncovers the time-series structure of outbreak severity for highly pathogenic avian influenza (H5N1) in Egypt.

III. OVERVIEW OF PROPOSED DESIGN

The main scope is the comparison of two machine-learning algorithms to obtain the accurate level of prediction model for the given data set. The overview of system design is depicted in Fig 1.

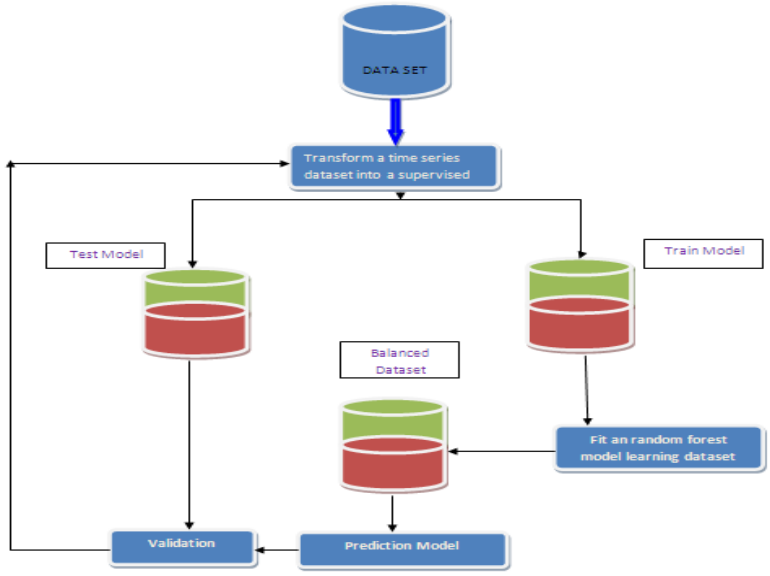


FIGURE 1.: Research methodology

IV. DATA COLLECTION

The data obtained on daily basis from the site https://mis.cept.gov.in/CBS/Dbt_Dash.aspx was in excel format collected circle wise for pan India ,the data was merged to single dataset and splitted according to small saving scheme products and date wise. Separate prediction were made for the each products using the models developed using ARIMA and Random forest, Data was transformed to Time series frame and data was splitted to train and test the model ,there by implementing the algorithm separately to develop the predictive model for each product of DOP and validation was carried out.

v. PRE-PROCESSING AND TRAINING

The collected data were then processed by using the Juypter Notebook using python 3.7.4in built anaconda application. We calculated and scaled the following the date times series and no of account opened at which they were observed. These measures were the training inputs of RF and ARIMA models. These models were trained with data from the last seven months of data obtained for each schemes, from previous to the desired date each of the RF and ARIMA model families was trained and optimized via 12-fold cross validation to acquire the prediction for next 12 months; however, this process is time consuming for real-time predictions with big data. Multiple and different models for each model family were tested on training results. The advantage of this process is that the model with the lowest Root Mean Square Error (RMSE) will be selected to predict average no of account opened from time series using ARIMA ,also with Mean Absolute Error in Random Forest Regression (MAE) will be selected to predict average no of account opened from time series using Random Forest .Different layers were tested and the model with the best accuracy was

selected to predict the future values of account opening for each schemes. This training process was applied to all models.

Running the example reports the expected and predicted values for each step in the test set, then the MAE for all predicted values. Note: Your results may vary given the stochastic nature of the algorithm or evaluation procedure, or differences in numerical precision. Consider running the example a few times and compare the average outcome. We can see that the model performs better than a persistence model, achieving a MAE.

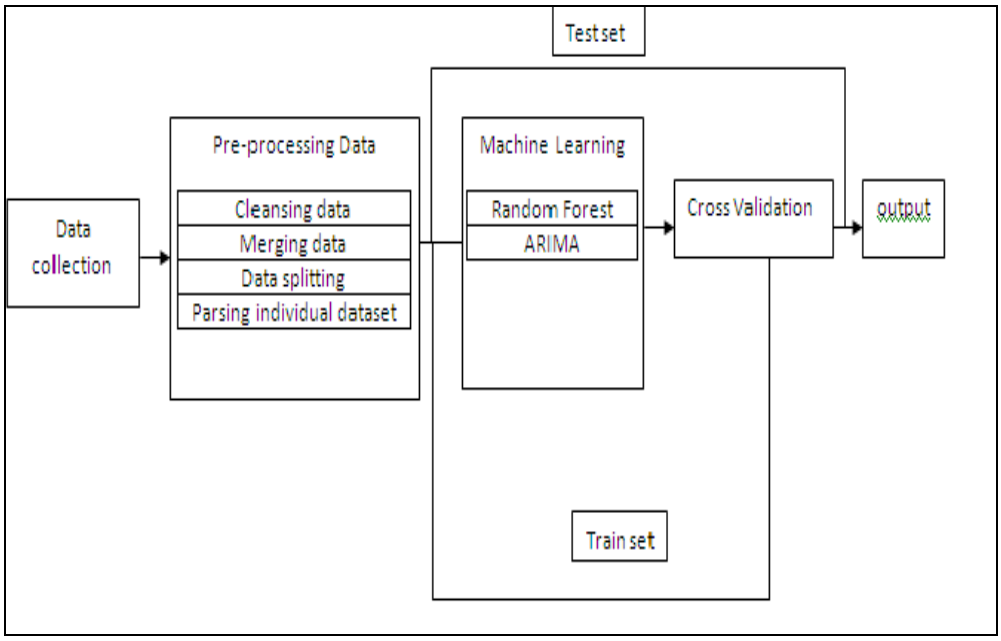


FIGURE 2. Pre-Processing and Training

VI. RANDOM FOREST MODEL

A Random Forest (RF) is an ensemble technique which can perform regression tasks by growing multiple decision trees and combining them with Bootstrap Aggregation, broadly known as bagging. First, the number of trees, ntree and bootstrap samples are drawn from the training set data. For each sample, a regression tree is fully grown, with the modification that at each node, rather than choosing the best split among all predictors, a random sample m try of the predictors is chosen and the best split is determined from those variables. The prediction of the mean speed is obtained by aggregating the predictions of the n tree trees, i.e., averaging the predictions[9s] .we can evaluate the Random Forest model on the dataset when making one-step forecasts for the last 12 months of data. We will use only the previous seven time steps as input to the model and default model hyper parameters, except we will use 1,000 trees in the ensemble (to avoid under learning).

vii. ARIMA MODEL

The stats models library provides the capability to fit an ARIMA model. An ARIMA model can be created using the stats models library as follows: Define the model by calling ARIMA and passing in the p, d, and q parameters. The model is prepared on the training data by calling

the fit() function. Predictions can be made by calling the **predict()** function and specifying the index of the time or times to be predicted. Let's start off with something simple. We will fit an ARIMA model to the entire Shampoo Sales dataset and review the residual errors.
First, we fit an ARIMA(5,1,0) model[11]. This sets the lag value to 5 for auto regression, uses a difference order of 1 to make the time series stationary, and uses a moving average model of 0.The ARIMA model can be used to forecast future time steps. We can use the predict function on the ARIMA Results object to make predictions. It accepts the index of the time steps to make predictions as arguments. These indexes are relative to the start of the training dataset used to make predictions. If we used 100 observations in the training dataset to fit the model, then the index of the next time step for making a prediction would be specified to the prediction function as start=101, end=101. This would return an array with one element containing the prediction.

Alternately, we can avoid all of these specifications by using the forecast function, which performs a one-step forecast using the model. We can split the training dataset into train and test sets, use the train set to fit the model, and generate a prediction for each element on the test set. A rolling forecast is required given the dependence on observations in prior time steps for differencing and the AR model. A crude way to perform this rolling forecast is to re-create the ARIMA model after each new observation is received. We manually keep track of all observations in a list called history that is seeded with the training data and to which new observations are appended each iteration. Putting this all together, below is an example of a rolling forecast with the ARIMA model in Python. We can also calculate a final root mean squared error score (RMSE) for the predictions.

VIII. RESULTS & DISCUSSION

The effectiveness of the algorithms' forecasting abilities was assessed via multiple comparisons. RMSE: Root Mean Squared Error and MAE: Mean Absolute Error are obtained,- The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. The RMSE will always be larger or equal to the MAE, the greater difference between them, the greater the variance in the individual errors , Both the MAE and RMSE can range from 0 to ∞. Lower values are better. Measuring Prediction Accuracy In order to assess the quality of the predictions, it is essential to establish metrics that allow the comparison of the different methods. This evaluation must consist of a comparison between the prediction results and the expected conditions at the selected date and total no of account opened. We used the following metrics:
Mean Absolute Error (MAE)—This metric corresponds to the average absolute difference between the predicted yˆ and the true values y as denoted in equation(1).

$$MAE = 1/n \sum_{j=1}^{n} |y_j - \hat{y_j}| \text{-------} (1)$$

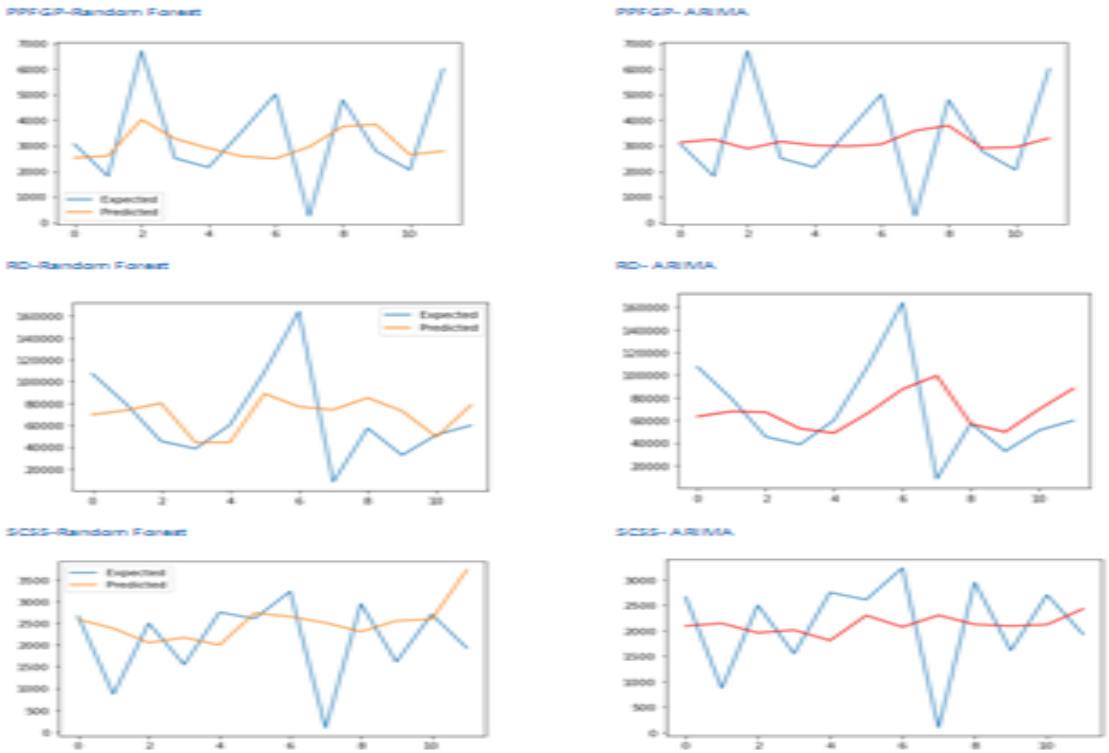
• Root Mean Square Error (RMSE)—This metric corresponds to the square root of the mean of the squared difference between the observed y and the predicted values yˆ y as denoted in equation(2)..

$$RMSE = \sqrt{1/n \sum_{j=1}^{n} (y_j - \hat{y_j})^2} \text{------} (2)$$

Table 1 MAE and RMSE of the models' predictions

MODEL/ PRODUCT	ARIMA		RANDOM FOREST	
	MAE	RMSE	MAE	RMSE
KVN	2449.367	3269.2	2635.538	3663.7
MIS	2361.542	2927.3	2427.93	3120.7
NSC8	4376.735	6149.4	4137.216	5504.2
PPFGP	1459.079	1885.5	1479.976	1774.7
RD	31360.9	41039.1	29413.36	38171.4
SBSGP	5857.31	8009.4	5626.735	8490.8
SCSS	822.817	965.8	829.764	1084.5
SSA	6456.251	7923	6940.582	8875.1
TD	8012.076	10776.2	8174.451	11101.4

ARIMA showed better performance with the minimum MAE and RMSE values and the minimum of maximum errors, as shown in Table 1.



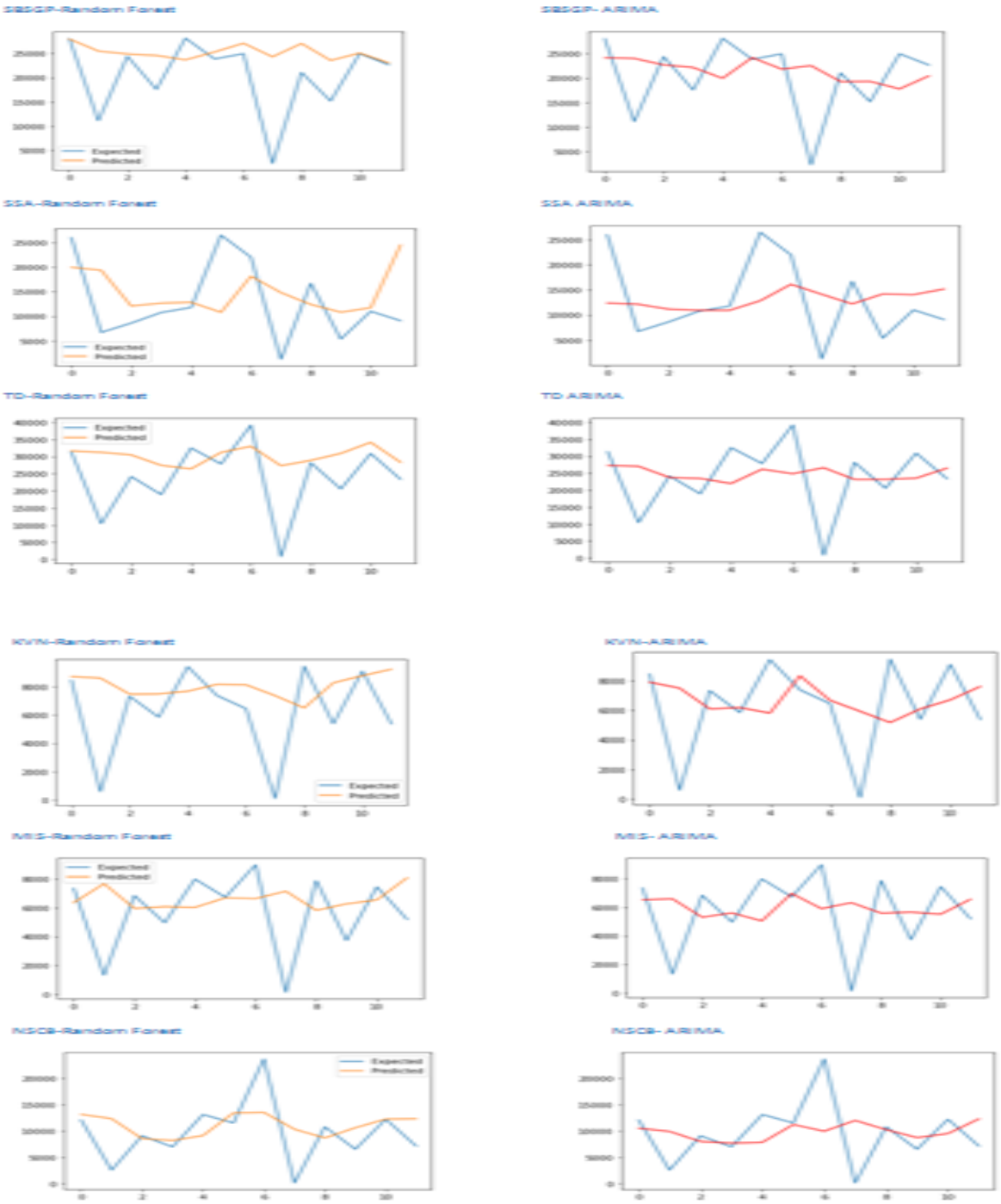


Figure 3. Predictions and Expected account opening values of products at twelve consecutive predictions next twelve months.

IX. CONCLUSIONS

Performing predictions was an effective means for handling distribution changes in the prospective setting. Assuming that the distribution changes occur slowly over time allows us to assume the process and the models can be trained to filter out. As a result of this windowing the ARIMA model likely outperformed the Random Forest model for two distinct reasons. First, the relationship between the expected and the predicted values are linear. From

Figure 3, where the importance for each of the variable used by the Random Forest model is shown expected is one of the most important variables used by the model to predict. However no corresponding term appears in the Random Forest model, which assumes linear relationships between the predicted and expected values. The ARIMA's has the ability to corporate with linear relationships may have contributed to better performance..Thus, we summarize three important aspects, including data preprocessing, data inputs, and evaluation rules. This study contributes to the literature by presenting a valuable accumulation of knowledge on related studies and providing useful recommendations for financial analysts and researchers. ARIMA model accommodates these dynamic relationships, updating the model based on recent events to predict the future state of the system.

X. FUTURE WORKS

My future work resides on extracting the data for Postal Life Insurance and Logistics services render by the Postal department and use the predictive algorithm which are suitable for the sector studies in future ,also to study the algorithm bench marking comparative studies with other compete organization.

XI. ACKNOWLEDGEMENT

This research was possible thanks to the data provided for research purposes by Department of Post's .

REFERENCES

1 Pavlyshenko, B. (2019) 'Machine-Learning Models for Sales Time Series Forecasting', Data, 4(1), p. 15. doi: 10.3390/data4010015.

2 Elias, S. and Singh, S. (2018) 'FORECASTING of WALMART SALES using MACHINE LEARNING ALGORITHMS'

3 Li, M., Ji, S. and Liu, G. (2018) 'Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model', Mathematical Problems in Engineering, 2018, pp. 1–12. doi: 10.1155/2018/6924960.

4 Bohanec, M., Kljajić Borštnar, M. and Robnik-Šikonja, M. (2017) 'Explaining machine learning models in sales predictions', Expert Systems with Applications, 71(April), pp. 416–428. doi: 10.1016/j.eswa.2016.11.010.

5 Mohammed, M., Khan, M. B. and Bashie, E. B. M. (2017) Machine learning: Algorithms and applications, Machine Learning: Algorithms and Applications. doi: 10.1201/9781315371658.

6 Pavlyshenko, B. (2019) 'Machine-Learning Models for Sales Time Series Forecasting', Data, 4(1), p. 15. doi: 10.3390/data4010015.

7 E-Commerce System for Sale Prediction Using Machine Learning Technique To cite this article: Karandeep Singh et al 2020 J. Phys.: Conf. Ser. 1712 012042.

8 Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks Michael J Kane, Natalie Price, Matthew Scotch & Peter Rabinowitz BMC Bioinformatics volume 15, Article number: 276 (2014) .

9 Random Forest for Time Series Forecasting by Jason Brownlee on November2,2 020 in Time Series[https://machinelearningmastery.com/random-forest-for-time-series-forecasting/]

10 Chen, H.; Rakha, H. Real-time travel time prediction using particle filtering with a non-explicit state-transition model. Transp. Res. Part Emerg. Technol. 2014, 43. [Cross Ref]

11 How to Create an ARIMA Model for Time Series Forecasting in Python by Jason Brownlee on January 9, 2017 in Time Series [https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/].