2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)
Bangalore, India. Oct 7-9, 2022

Vaccine Supply Optimization and Forecasting using Random Forest and ARIMA Models

Shiv Charan Banerjee
Department of Engineering Technology
Birla Institute of Technology and Science - Pilani
Rajasthan, India
shiv.banerjee@nic.in

Shobhan Banerjee
Graduate Student Member IEEE
Department of Engineering Technology
Birla Institute of Technology and Science - Pilani
Rajasthan, India
shobhanbanerjee3@gmail.com

Pratik Rai
Department of Operations Management
Indian Institute of Management Ranchi
Jharkhand, India
pratik.rai19ph@iimranchi.ac.in

2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT) | 978-1-6654-6855-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCAT55367.2022.9972154

***Abstract* — In order to tackle the Corona Virus Disease, it took a considerable amount of time for the governments to come up with effective and efficient vaccines. After the vaccines were developed, the next challenge was to supply the vaccines to various designated centers based on demographics, population distribution, and other factors. The whole system for vaccine supply played a vital role during the COVID-19 pandemic. We also saw a lot of haphazard and mismanagement in some places especially when the cases per day surged high, as people weren't prepared for such a situation. Now that we have got enough data, we can use it to optimize the vaccine supply across various Covid Vaccination Centers and be prepared for any such circumstances in the future. In this paper, we have proposed a two-step approach where considering the past supply and wastage data we performed a classification task that indicates whether doses are to get wasted at a given center. If yes, we then perform demand forecasting based on the number of administered doses so that the wastage can be reduced, and supply can be optimized.**

Keywords—Classification, Forecasting, Supply Optimization, Random Forest, ARIMA

I. Introduction

Centuries ago, since transportation was not much developed hence the wide-scale spread of diseases used to get limited to endemics or epidemics. The H1N1 influenza pandemic which happened after World War – 1, spread gradually when soldiers were returning to their homes. Due to globalization, since transportation has developed these days, we saw a rapid widespread of the COVID-19 pandemic across the globe.

When it comes to the supply and management of vaccines, the task becomes quite challenging as it technically takes the form of multi-variate analysis based on which one has to optimize the whole supply chain. This task becomes more challenging when the dose has to be administered in phases (e.g., the first dose, second dose, precautionary dose, etc.). The demand for vaccines varies with respect to demographics. There are certain other parameters as well that need to be considered for optimization of the whole vaccine supply including overdosage, under dosage, wastage, etc.

For example, from a given vial (unit quantity) fewer or more people can receive the dosage than it was meant to be. While considering the population distribution, we might have to consider the children below 15 years of age separately. The same thing goes for adults over 60 years of age. Or sometimes the days of the week, holidays, etc. might play a deciding role in an increase or decrease in the vaccine requirements.

An intelligent distribution system based on economic impacts has been proposed by authors in [1] using the data for the city of Jordan. Authors in [2] have presented a blockchain-based vaccination scheme that ensures data integrity and immutable registrations, hence avoiding identity thefts. In [3] the authors have proposed a blockchain and ML-based approach that monitors the prerequisites of vaccine distribution along with demand forecasting. The authors have worked on immunization in [4], for managing and tracing stocks and logistics to ensure transparent distribution. An Ethereum blockchain-based solution has been proposed by authors in [5] to manage data related to vaccine distribution and delivery.

Authors in [6] have proposed a decision support system integrating semaphores to facilitate the distribution and administration processes of CTRI/2020/08/027170 and CTRI/2020/11/028976. Authors in [7] have discussed a security framework to distribute vaccines and track them using an IoMT-based environment. VaCoChain has been proposed in [8] that integrates blockchain and UAVs for timely vaccine distribution. The UAV application proposed in [9] can also be used for accurate target detection and delivery of vaccines of any sort. In [10], the authors have explored how the supply chain applications of blockchain impact workflow among organizations by modeling the vaccine distributions. VaxEquity – an optimization framework has been proposed by authors in [11] that performs data-driven risk assessment for equitable vaccine distribution.

In this paper, we have proposed a method that involves two steps. In the first step, we used a machine learning algorithm to predict whether a Covid Vaccination Centre (CVC) will have a wastage in vaccination dosage. If there is no wastage, then there's no need to consider the CVC. If at all a wastage is expected to occur at a given CVC on a given day, then we run a probabilistic time-series analysis model which forecasts the expected number of doses to be administered based on the past 3 months of vaccine supply data in that CVC.

For classification in the first step, we used a random forest classifier with optimal hyperparameter tuning followed by an ARIMA model for the second step.

II. The Dataset

The data used for this work have been collected from the District Administration Ranchi, Civil Surgeon Office, Office of District Immunization Officer and Epidemiologist, and Integrated Disease Surveillance Program (IDSP) after taking all the necessary and required permissions from the District Administration.

The data was available to us in a scattered form, which was initially extracted from all the relevant sources. Followed by which this data was transformed and merged to meet our requirements. Finally, we loaded the data and proceeded ahead with modeling. We came up with nine important attributes namely: CVC, Rurality, Comorbidity, Wastage, #HCW, #FLW, Age Group, #female participants, and Day of the Week which have been discussed below.

1. *CVC*:

   This indicates the name of the Covid Vaccination Centre.

2. *Rurality*:

   This indicates whether the CVC is in a rural or urban region.

3. *Comorbidity*:

   This indicates the number of people in the immune-comprised age group, i.e., they are suffering from blood pressure, diabetes, etc.

4. *Wastage*:

   The state's average threshold is 5% net wastage. Below 5%, wastage is not considered. This has been used as our target attribute.

5. *#HCW*:

   These are the counts of the number of Health Care Workers such as doctors, nurses, Emergency Medical Technicians (EMTs), paramedical staff, vaccinators, etc.

6. *#FLW*:

   These include the Front-Line Workers such as police personnel, magistrates, government officials, municipal corporations, etc.

7. *Age Group*:

   This consists of various age ranges, i.e., 15-17, 18-44, 45-59, and 60 above. The age range below 15 years has not been considered in this study.

8. *#Female Participants*:

   This has been used because, in rural areas, females are occupied all throughout the day and hence are less aware and responsive towards vaccination, which is one of the major causes of wastage.

9. *Day of the week*:

   This parameter has been considered as the number of people turning up weekends or holidays might be more if compared to that on weekdays.

Considering the values of the target attribute above, based on the threshold value mentioned above, the objective took the form of a binary classification problem with label values - 1 and 0. After removing all the duplicate values and imputing the missing values, we got a dataset of dimensions 173×9.

III. Random Forest Classification

Considering the structure of the data available to us, the requirement was to go with a model which is:

1. Non-parametric – no assumptions about the dataset's shape
2. Capable to work with categorical & numerical features.
3. Requires minimum data preprocessing
4. Feature selection happens by itself and unimportant features won't influence the output.
5. Multicollinearity doesn't affect the quality of classification.

Keeping the above points in consideration, tree-based models work best in such scenarios. But since they tend to overfit, we decided to address the issue using in two ways:

1. Using an ensemble model for classification (Random Forest Classifier).
2. Performing a Grid Search Cross-validation to find the optimal hyperparameters and fine-tune our model to obtain the best results.

For Cross Validation, the number of decision-making trees was kept as – [20, 30, 50, 100, 200], the maximum depth of the trees was kept as – [3, 5, 10], and the minimum samples for a leaf node as – [5, 10, 20]. All the processors were used in parallel in our classifier. After performing a 5-fold cross-validation, the best estimators turned up to be 200 trees at a maximum depth of 3 with a minimum number of samples in a leaf node being 20.

The training and test sets were split in the ratio of 70:30, hence having 121 instances for training and cross-validation, and the rest 52 for testing the accuracy of the model. The top 5 features and their importance values are shown in the figure below:

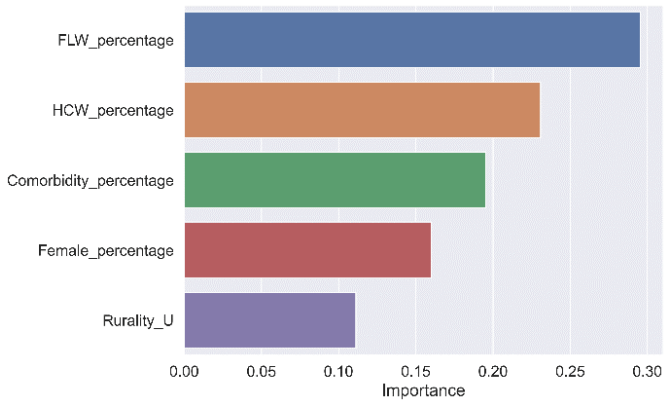

Fig.-1: Top 5 features and their importance

The cumulative sums of the feature importances add up to 100 in the figure above.

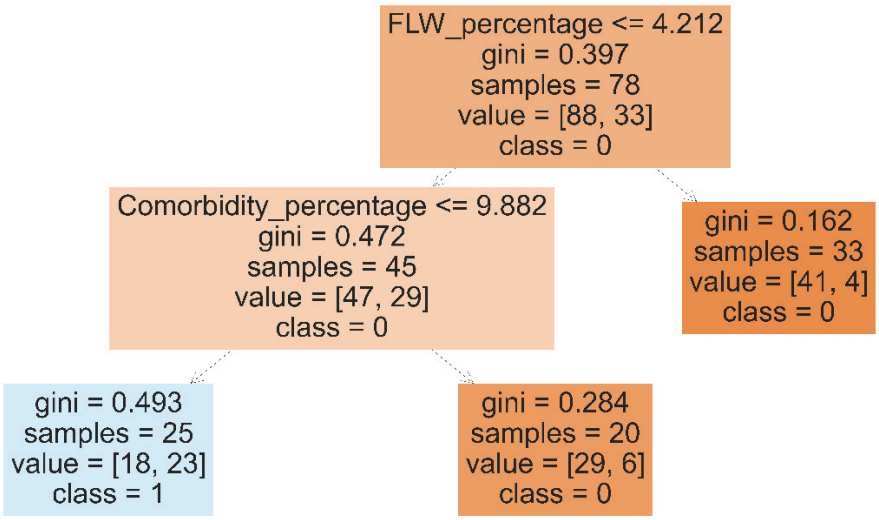

Fig. 2: One of the best Trees from the Forest

Figure 2 above shows the structure of one of the best trees that we get after using the best estimators. Figure 3 below shows the Receiver-Operating Characteristics for the best-acquired model, with an area of 81% lying under the curve.

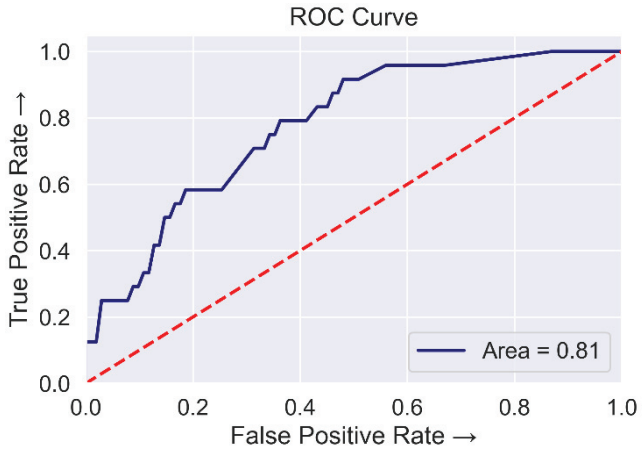

Fig.3: Receiver Operating Characteristics

The model gave us a classification accuracy of 94.23%, the confusion matrix for which is shown in the figure below.

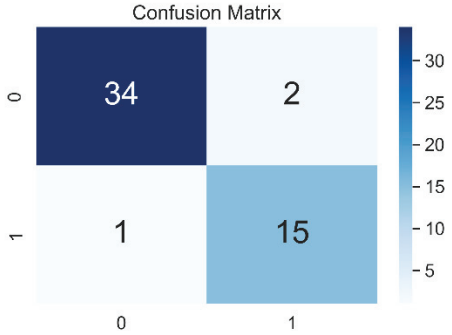

Fig. 4: The Confusion Matrix

IV. ARIMA Model

It stands for Auto-Regressive Integrated Moving Average model. Whenever we intend to perform univariate time-series forecasting, ARIMA Model is one of the suitable linear regression models. From the previous step as explained in Section II, if we get to have wastage, then we find the CVCs corresponding to them (which are almost common), followed by which ARIMA model has been used on the past three months' data for April, May, and June of that CVC to predict the future requirements of dosage to be administered.

The result for one of such major CVC – RISALDAR BABA has been shown for illustration The plot for the number of administered doses is shown in the figure below.

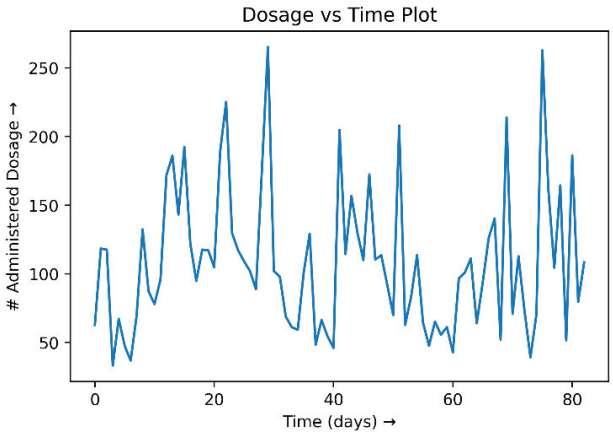

Fig. 5: Administered dosage for Risaldar Baba

The spikes on certain days show that the number of supplied doses and the number of administered doses were the same on certain days, but on other days the difference is too high, which means that there has been a significant amount of wastage.

Keeping the significance level as 0.05 and stating the null and alternate hypotheses as:

P_0: Time series is non-stationary

P_A: Time series is stationary

After performing Augmented Dickey-Fuller Test, we acquired a P-value of 0.0043. Hence, P_0 gets rejected indicating the statistical significance of the test, and no further differencing is needed, i.e., d=0.

Next, we need to identify if the model needs an AR term. Hence, we inspect the Partial Auto-Correlation Factor (PACF). We see that the PACF lag 1 is significant, and the rest are well within the significance region, hence we take p=1. The PACF plot has been shown in the figure below:

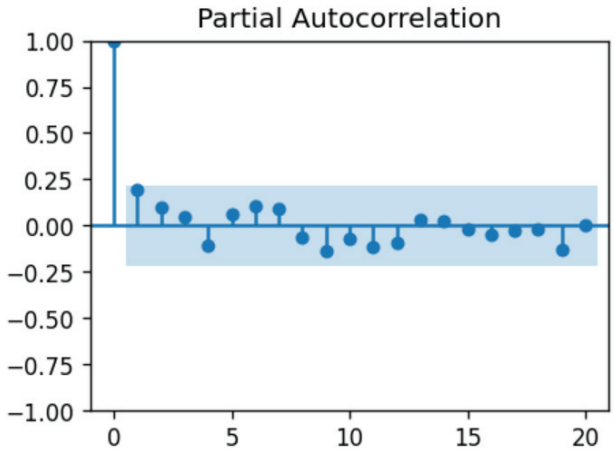

Fig. 6: PACF Plot

Next, we need to identify if the model needs an MA term. Hence, we inspect the Auto-Correlation Factor (ACF). We see that in this case as well, the ACF lag 1 is significant, and the rest are well within the significance region, hence we take q=1. The ACF plot has been shown in the figure below:

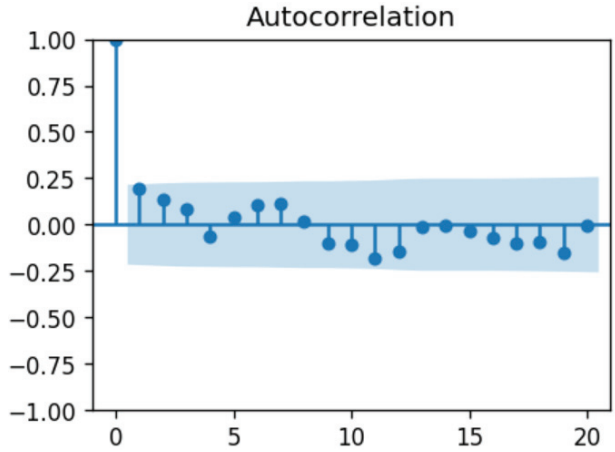

Fig. 7: ACF Plot

Now that we have p=1, q=1, and d=0, we proceed ahead with our ARIMA modeling. The model gave us an AIC_1 = 495.663

We tried another combination with p=1, q=0, and d=0 which yielded slightly better AIC_2 = 494.447.

V. Results and Discussion

The plots for residual error and density have been shown in figures 8 and 9 respectively. We can infer from the residuals plot that the error in regression lies within the range +10 to -5.

We chose Root Mean Square Error (RMSE) as the accuracy metric to evaluate our model where

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Xact - Xpred)^2}{N}}$$

The RMSE value turned out to be 3.388 for our model. This means that if one vial of dosage is considered a bias, that would take care of the requirements for RISALDAR BABA CVC.

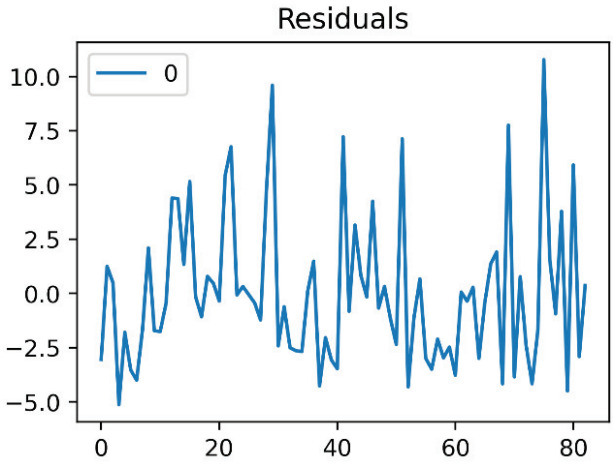

Fig. 8: Residuals Plot

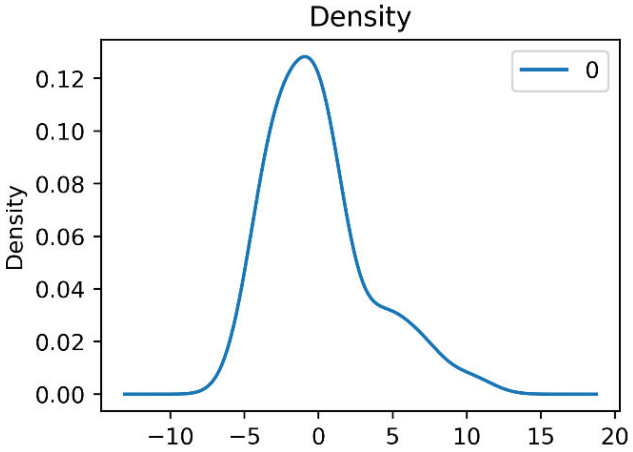

Fig. 9: Density Plot

VI. Conclusion & Future Scope

From the density plot, as shown above, in figure 9, we can clearly see that the errors are mostly concentrated toward zero. However, we can certainly try to improve our accuracy by using other such probabilistic models or even deterministic models where we can get more optimized results. The same analysis can be extended to all other CVCs where we get to have wastage from Step 1 in our work and proceed with analysis at a granular level for all such CVCs ensuring minimum wastage and optimal supply.

References

[1] S. AlZu'bi, Q. H. Makki, Y. A. Ghani and H. Ali, "Intelligent Distribution for COVID-19 Vaccine Based on Economical Impacts," 2021 International Conference on Information Technology (ICIT), 2021, pp. 968-973, doi: 10.1109/ICIT52682.2021.9491787.

[2] C. Antal, T. Cioara, M. Antal and I. Anghel, "Blockchain Platform For COVID-19 Vaccine Supply Management," in IEEE Open Journal of the Computer Society, vol. 2, pp. 164-178, 2021, doi: 10.1109/OJCS.2021.3067450.

[3] T. I. Meghla, M. M. Rahman, A. A. Biswas, J. T. Hossain and T. Khatun, "Supply Chain Management with Demand Forecasting of Covid-19 Vaccine using Blockchain and Machine Learning," 2021 12th International Conference on Computing Communication and

Networking Technologies (ICCCNT), 2021, pp. 01-07, doi: 10.1109/ICCCNT51525.2021.9580006.

[4] A. V, D. J. S, D. V. A and V. Krishna Kumar, "Blockchain Based Covid Vaccine Booking and Vaccine Management System," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1-7, doi: 10.1109/ICOSEC51865.2021.9591965.

[5] A. Musamih, R. Jayaraman, K. Salah, H. R. Hasan, I. Yaqoob and Y. Al-Hammadi, "Blockchain-Based Solution for Distribution and Delivery of COVID-19 Vaccines," in IEEE Access, vol. 9, pp. 71372-71387, 2021, doi: 10.1109/ACCESS.2021.3079197.

[6] A. K. N. S. M. Gayathri, A. R. Pai and J. S. R. R, "A Decision Support System to facilitate Vaccination for Covid-19 pandemic," 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022, pp. 1-4, doi: 10.1109/ICCCI54379.2022.9740797.

[7] A. K. Das, B. Bera and D. Giri, "AI and Blockchain-Based Cloud-Assisted Secure Vaccine Distribution and Tracking in IoMT-Enabled COVID-19 Environment," in IEEE Internet of Things Magazine, vol. 4, no. 2, pp. 26-32, June 2021, doi: 10.1109/IOTM.0001.2100016.

[8] A. Verma, P. Bhattacharya, M. Zuhair, S. Tanwar and N. Kumar, "VaCoChain: Blockchain-Based 5G-Assisted UAV Vaccine Distribution Scheme for Future Pandemics," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 5, pp. 1997-2007, May 2022, doi: 10.1109/JBHI.2021.3103404.

[9] T. Swain, M. Rath, J. Mishra, S. Banerjee and T. Samant, "Deep Reinforcement Learning based Target Detection for Unmanned Aerial Vehicle," 2022 IEEE India Council International Subsections Conference (INDISCON), 2022, pp. 1-5, doi: 10.1109/INDISCON54605.2022.9862891.

[10] H. M. Chung, "Blockchain-based Value Creation for Supply Chain in COVID-19 Drug Distribution," 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), 2022, pp. 1-6, doi: 10.23919/CISTI54924.2022.9820529.

[11] N. Kaur, J. Hughes and J. Chen, "VaxEquity: A Data-Driven Risk Assessment and Optimization Framework for Equitable Vaccine Distribution," 2022 56th Annual Conference on Information Sciences and Systems (CISS), 2022, pp. 25-30, doi: 10.1109/CISS53076.2022.9751173.