JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1875 - 1882
https://publishoa.com
ISSN: 1309-3452

Machine learning approach on Indian healthcare consumer data

Challa Maruthy Subrahmanyam[1], Dr Sarita Satpathy[2], Dr S. K Satpathy[3]

[1] Research Scholar, Department of Management Studies, Vignan's Foundation for Science, Technology and Research University, Guntur, A.P, India– 522213 , E mail: maruthychalla@yahoo.com
[2] Associate Professor, Department of Management Studies, Vignan's Foundation for Science, Technology and Research University, Guntur, A.P, India– 522213, E mail: ssssatpathy3@gmail.com
[3]Associate Professor, Department of CSE, Vignan's Foundation for Science, Technology and Research University, Guntur, A.P, India– 522213, E mail: drsks_cse@vignan.ac.in

ABSTRACT
Revisiting of an outpatient to hospital is an important parameter to evaluate the performance of a hospital. It depends on the outpatient perceptions of healthcare services by a hospital. In this study 416 outpatient's perception is captured through an online questionnaire. Machine learning approach is used to predict the revisiting outpatient to hospital. Ten machine learning algorithm (MLA) were applied on the healthcare consumer data. Top 5 and bottom 5 features on the basis of feature importance is derived from the MLA. This can be used as indicator by hospital administrator to take remedial actions when the revisiting outpatient's number is below the desired level. The best MLA in terms of accuracy is SVC.

Keywords: Outpatient, machine learning, revisiting, hospital

1. Introduction

Out-patient performs following activities: searches for healthcare provider collects data on the healthcare provider, searches on the medical specialist, books an appointment or walk in at the healthcare provider, physically reaches the healthcare provider, meets the reception, waits for the medical specialist, medical specialist gives prescription to the healthcare consumer, buys medicines, takes diagnostic tests and probably visits canteen. Every touch point of the outpatient with the healthcare provider leaves a perception with the outpatient. Outpatient's perception is the sum of all the individual perception gained from all the touch points. This affects outpatient's decision in selecting a hospital.

Hospital would like all their outpatients to revisit when need arises. A predicting model for 'revisiting of an outpatient' will indicate the performance of the hospital. The hospital administrator can take appropriate action accordingly.

1.1 Literature survey

Machine learning is used in many fields. Healthcare also uses machine learning (Chen, Pierson, et al., 2021). Diagnosis is one of the fields in healthcare where machine learning can be used to make healthcare better by adopting the machine learning algorithms (Gerard et al., 2008). The other field of healthcare where machine learning can be used is in treatment of diseases, remote patient monitoring (Saleem & Chishti, 2019) and in patient documents (Toh & P. Brody, 2021). Machine leaning can give the complete picture of the healthcare data (Chen, Joshi, et al., 2021).

This paper tries to find the best machine learning algorithms to arrive the healthcare consumer to revisit the healthcare facility in need. Some of the earlier work done are :

Kyun Jick Lee in his research paper 'A Practical Method of Predicting Client Revisit Intention in a Hospital Setting' states the five important predictors of revisit intention as overall satisfaction, intention to recommend to others, awareness of hospital promotion, satisfaction with physician's kindness, and satisfaction with treatment level. This is based on a study in South Korea (Lee, 2005).

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1875 - 1882
https://publishoa.com
ISSN: 1309-3452

Singh & Goyal in their research paper to find the best machine learning algorithm for cervical cancer concluded logistic regression is best suited for the cervical data set (Singh & Goyal, 2020). Dhwaani Parikh and Vineet Menon in their research paper 'Machine Learning Applied to Cervical Cancer Data' have chosen the best performing machine learning algorithm K-nearest neighbor amongst decision tree and random forest algorithm (Dhwaani & Vineet, 2019).

Sandhya in research paper presented a machine learning model for hospital recommendation which they were implementing in a application (B, 2020). Mani research paper presents a software based on machine learning the suggests the names of nearby hospital basing on patients disease condition (V, 2020).

2. Methodology

2.1 Data acquisition

An online questionnaire developed in Google form is used to collected data from outpatients .The questionnaire captured the outpatient's perception during his visit to the hospital. The following Table 1 gives the description of the data acquired.

Table 1: Data description

S.No	Column	Description
1	Gender	Gender of respondent
2	Age	Age of respondent
3	Income	Income of respondent
4	Advertisement	Advertisement done by healthcare provider
5	Choose	Choose healthcare provide for
6	Rdoctor	Read on the medical specialist
7	Appointment	Appointment made using
8	Clean	Premises were clean and hygienic
9	Softused	Software used in making appointment
10	Insuranceco	Medical insurance is from
11	Equipment	Medical equipment
12	Waitingarea	Waiting area for out-patient
13	Canteen	Hospital canteen
14	Opreseception	Out-patient reception
15	Feedback	Feedback collection system
16	Confidence	Doctor while treating exhibited confidence
17	Testexplain	Doctor explains the diagnostic tests
18	Rateservices	Rate serviced of hospital
19	Staffcoop	Hospital staff cooperation
20	Registrationtime	Time taken for out-patient registration
21	Waitingtime	Waiting time for doctors
22	Purchasingtime	Time to purchase medicines
23	Timereports	Time taken to collect reports
24	Hosppresciption	Doctors prescription
25	Informdiseases	Information on disease
26	Feedbackcustomer	Feedback by earlier consumers
27	Suggestpatient	Consumer suggestions
28	Pricingopinion	Opinion on hospital pricing
29	Hospcompetitor	Hospital competitor
30	Revisit	Revisit by healthcare consumer

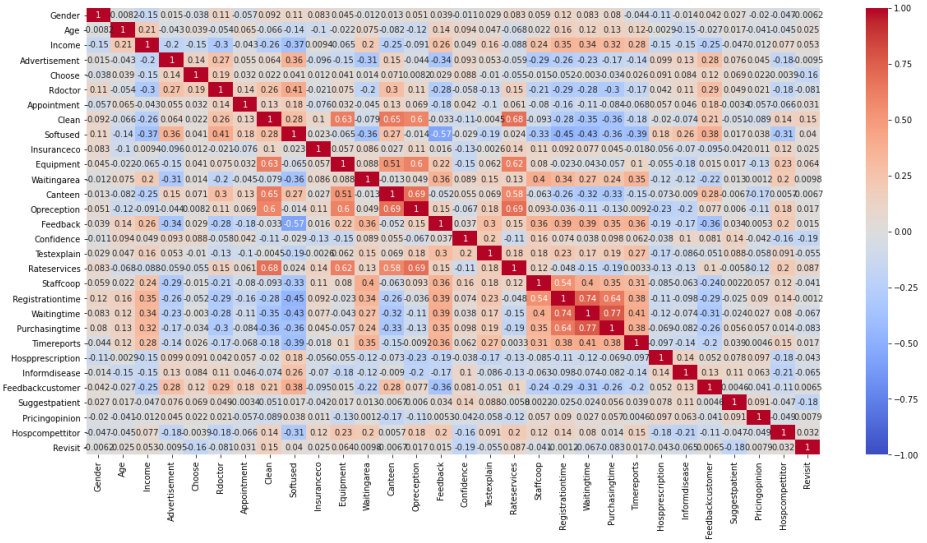
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1875 - 1882
https://publishoa.com
ISSN: 1309-3452

2.2 Data Exploration

The data has 416 rows and 30 columns. The output is in 'revisit' column. Remaining columns provides perception of the outpatient. All the columns contained integer data type (int64).

The correlation matrix between features is shown in Fig 1.

Figure 1: Correlation matrix between features



Two columns 'waitingtime' and 'rateservices' have high correlation values.

The frequency count of the 'Revisit' column shows number of rows containing 1 are 205 and 2 are 211.

2.3 Machine learning algorithm comparison

Ten machine language algorithms (MLA) are used to create the model for MLA comparison. The ten algorithms are

a. Logistic Regression
b. Linear Discriminant Analysis
c. SVC
d. Decision TreeClassifier
e. Random Forest Classifier
f. Gradient Boosting Classifier
g. Ada Boost Classifier
h. XGB Classifier
i. KNeighbors Classifier
j. LGBM Classifier

Steps involved in the machine learning algorithm comparison are

Step 1: Initial model building

The ten machine learning algorithms (MLA) are run with the data (rows=416, columns=30). They are sorted in descending order on the basis of accuracy as shown in Table.2.

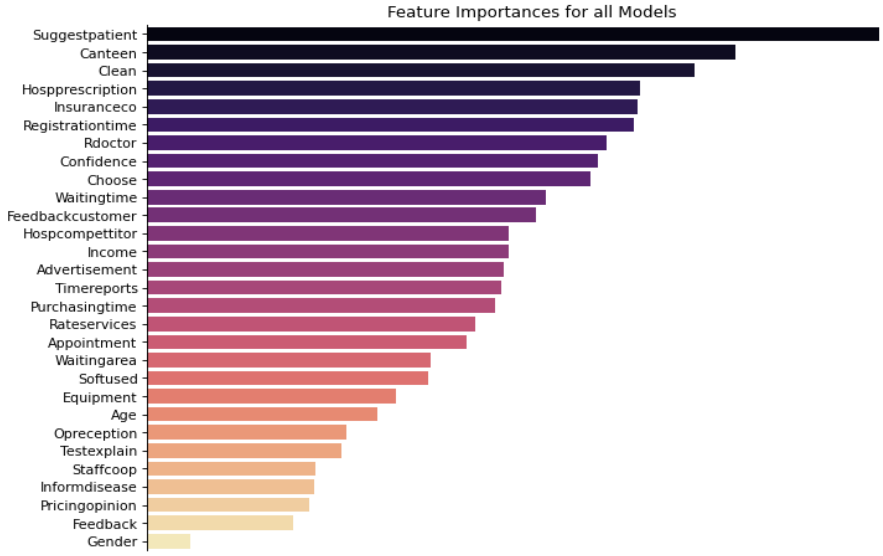
Table 2:Initial model accuracy

S.No	MLA name	Training accuracy mean	Test accuracy mean

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1875 - 1882
https://publishoa.com
ISSN: 1309-3452

1	LGBMClassifier	0.996787	0.6112	
2	DecisionTreeClassifier	1	0.6096	
3	SVC	0.804819	0.6	
4	XGBClassifier	0.959839	0.5968	
5	GradientBoostingClassifier	0.990361	0.5888	
6	RandomForestClassifier	0.987149	0.5872	
7	AdaBoostClassifier	0.803213	0.5872	
8	KNeighborsClassifier	0.726104		0.5856
9	LinearDiscriminantAnalysis	0.663454	0.5616	
10	LogisticRegression	0.665863	0.5568	

Step 2: Feature importance
Feature importance (see **Error! Reference source not found.**) from each algorithm is extracted and sorted in descending order. SVM and K – Neighbors do not give feature importance. Thus they are not included in the feature importance.
Figure 2: Feature Importance for all the models



.

Step 3: Standardize the data
Data is split as input features and output feature. The inputs features are standardized using StandardScalar(). Standardizing of the sample is represented by the equation 1 where x = sample, u = mean of the sample, s=standard deviation and z=standard score

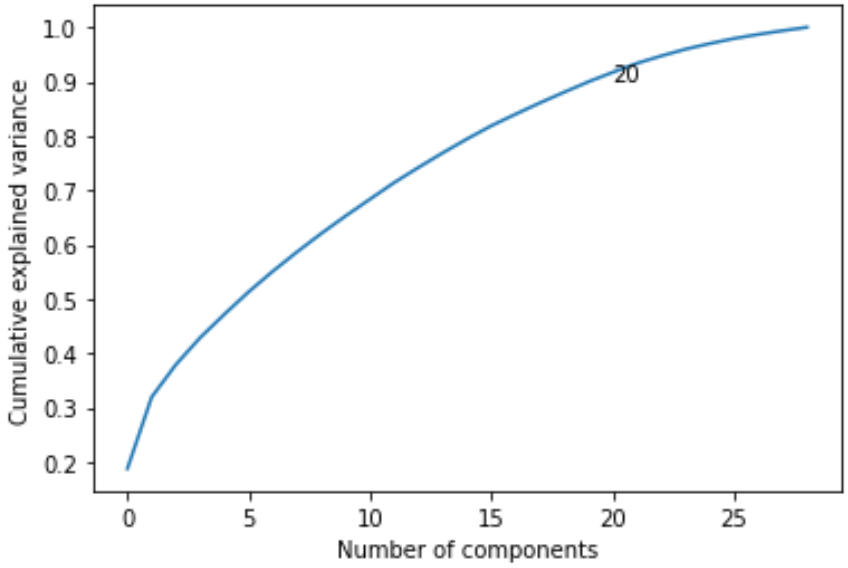
$$z = (x - u) / s \quad (1)$$

Step 4: Principal component analysis (PCA) on data
PCA is used on the data for dimension reduction and reduce correlations of features.

a. Dimension reduction
Fig 3 gives the graph of cumulative explained variance and number of PCA components. From the graph it is evident that 20 PCA components account for 92 % variance. These 20 PCA components are used in subsequent model building.

Figure 3: Graph between cumulative explained variance and number of PCA components

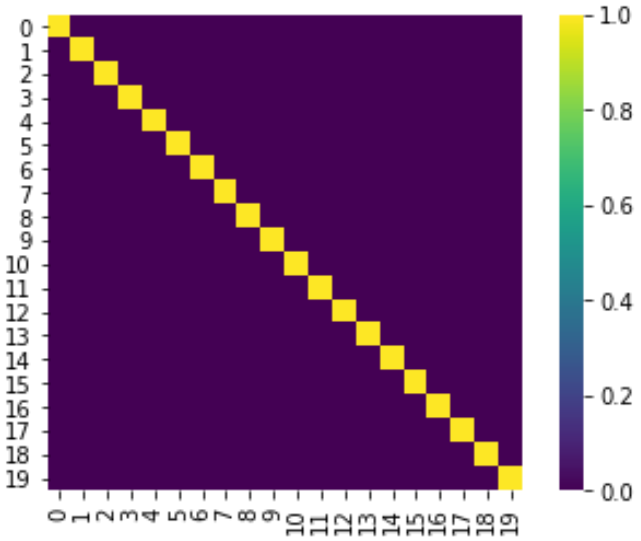
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1875 - 1882
https://publishoa.com
ISSN: 1309-3452



b. Reducing the correlation between input features

The correlation matrix between PCA features is given in Fig 4. From correlation matrix it is evident that there is reduction in correlation.

Figure 4: Correlation matrix between PCA components



Step 5: Model comparison for MLA after PCA

The ten machine learning algorithms (MLA) are run with the data (rows=416, columns=20). They are sorted in descending order on the basis of accuracy as shown in Table.3.

Table 3: Initial model for comparison of MLA after PCA accuracy

S.No	MLA name	Training accuracy mean	Test accuracy mean
1	SVC	0.82249	0.6176
2	KNeighborsClassifier	0.714859	0.6096
3	XGBClassifier	0.998394	0.5888
4	LGBMClassifier	1	0.5824
5	RandomForestClassifier	0.980723	0.5776
6	AdaBoostClassifier	0.93012	0.5712
7	GradientBoostingClassifier	1	0.5664

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1875 - 1882
https://publishoa.com
ISSN: 1309-3452

8	LinearDiscriminantAnalysis	0.635341	0.5536
9	LogisticRegression	0.629719	0.552
10	DecisionTreeClassifier	1	0.5184

3. Result and discussion

'Revisit' column shows number of rows containing 1 are 205 and 2 are 211. The difference between both the classes is 6.The data is assumed as balanced.

Many features have shown correlation more the 0.50 in correlation matrix. PCA is used to reduce the correlation.

The table 4 below shows top 5 and bottom 5 in feature importance. These can be indicator for the hospital administrator to take remedial actions if the revisiting outpatients are not to the desired number. The hospital administrator has to focus on the top 5 feature in case the number of revisiting outpatients is below the desired number.
Table 4: Top 5 and bottom 5 features

S.No	Top 5 features	Bottom 5 featires
1	Suggestpatient	Gender
2	Canteen	Feedback
3	Clean	Pricingopinion
4	Hosppprescription	Informdiseases
5	Insuranceco	Staffcoop

PCA gave 20 PCA components which account for 92 % variance. These are used for developing the model further.

The table 5 indicates the top 2 MLA initially and after PCA in terms of accuracy.
Table 5: MLA initial model and model after PCA

S.No	Initial Model	Model after PCA
1	LGBMClassifier	SVC
2	DecisionTreeClassifier	KNeighborsClassifier

4. Conclusion

Shape of initial data consists of 30 columns and 416 rows. In the initial model, MLA with highest accuracy is LGBM Cla ssifier. In feature importance the top feature is 'Suggestpatient'. The data is standardized. There are 20 PCA components which are taken for further model building. The MLA with highest accuracy after PCA is SVC.

References

1.Kyun Jick Lee, A Practical Method of Predicting Client Revisit Intention in a Hospital Setting, April 2005Health care management review 30(2):157-67
2.Ogbeyi Ofikwu Gabriel, Adekwu Amali and Amede Peter, Assessing the Level of Clients' Satisfaction on Outpatient and Inpatient Health Care Services, in a Tertiary Institution in North Central Nigeria, International Journal of Contemporary Medical Research, Volume 5 Issue 3 , March 2018, ICV: 77.83, ISSN (Online): 2393-915X; (Print): 2454-7379
3.Dhwaani Parikh and Vineet Menon, Machine Learning Applied to Cervical Cancer Data, I.J. Mathematical Sciences and Computing, 2019, 1, 53-64, Published Online January 2019 in MECS (http://www.mecs-press.net), DOI: 10.5815/ijmsc.2019.01.05, Available online at http://www.mecs-press.net/ijmsc
4.Juan M. Gorriz, Javier Ramirez, F. Segovia, Francisco J. Martinez, Meng, Chuan Lai, Michael V. Lombardo, Simon Baron-Cohen, Mrc Aims Consortium, John Suckling; A Machine Learning Approach to Reveal the Neuro-Phenotypes of Autisms; International Journal of Neural Systems; https://doi.org/10.1142/S0129065718500582 last accessed 5/5/2020

5. George D. Magoulas ,Andriana Prentza; Machine Learning in Medical Applications; Chapter · September 2001; DOI: 10.1007/3-540-44673-7_19 · Source: DBLP
6. Dhwaani Parikh, Vineet Menon; Machine Learning Applied to Cervical Cancer Data; I.J. Mathematical Sciences and Computing, 2019, 1, 53-64 Published Online January 2019 in ECS (http://www.mecs-press.net) DOI: 10.5815/ijmsc.2019.01.05; http://www.mecs-press.net/ijmsc last accessed 26/4/2020
7. M.A.Jabbar1, Shirina Samreen, Rajanikanth Aluvalu; TheFuture of Health care: Machine Learning; International Journal of Engineering & Technology; Website: www.sciencepubco.com/index.php/IJET last accessed 2/5/2020
8. Jenni A. M. Sidey-Gibbons and Chris J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, Sidey-Gibbons and Sidey-Gibbons BMC Medical Research Methodology (2019) 19:64
9. Ashish Khare,Moongu Jeon,Ishwar K. Sethi and Benlian Xu4, Machine Learning Theory and Applications for Healthcare, Journal of Healthcare Engineering, volume 2017, Article ID 5263570, 2 pages. https://doi.org/10.1155/2017/52635
10. Jenna Wiens and Erica S. Shenoy, Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology, Clinical Infectious Diseases, Volume 66, Issue 1, 1 January 2018, Pages 149–153, https://doi.org/10.1093/cid/cix731, Published: 21 August 2017
11. Sumeet DuaU, Rajendra Acharya and Prerna Dua, Machine Learning in Healthcare Informatics, Part of the Intelligent Systems Reference Library book series (ISRL, volume 56), Print ISBN 978-3-642-40016-2, Series Print ISSN 1868-4394
12. Arwinder Dhillon and Ashima Singh, Machine Learning in Healthcare Data Analysis: A Survey, Journal of biology and Today's World, http://journals.lexispublisher.com/jbtw, dol: 10.15412/J.JBTW.01070206
13. Ashish Khare, Moongu Jeon,Ishwar K. Sethi and Benlian Xu, Machine Learning Theory and Applications for Healthcare, Journal of Healthcare Engineering Volume 2017, Article ID 5263570, 2 pages, https://doi.org/10.1155/2017/5263570
14. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future healthcare journal, 6(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94
15. J. Ashok Kumar and S. Abirami, AN EXPERIMENTAL STUDY OF FEATURE EXTRACTION TECHNIQUES IN OPINION MINING, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.4, No.1, February 2015
16. Sandhya, Monika K. J , Vyshnavi and Usman K, Disease Prediction and Hospital Recommendation using Machine Learning Algorithms, International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue V May 2020- Available at www.ijraset.com
17. Abdullah R, Fakieh B. Health Care Employees' Perceptions of the Use of Artificial Intelligence Applications: Survey Study. Journal of Medical Internet Research 2020;22(5):e17620
18. Van Hartskamp M, Consoli S, Verhaegh W, Petkovic M and van de Stolpe A, Artificial Intelligence in Clinical Health Care Applications: Viewpoint, Interact J Med Res 2019;8(2):e12100, DOI: 10.2196/12100, PMID: 30950806, PMCID: 647320
19. Beaulieu-Jones B, Finlayson SG, Chivers C, et al. Trends and Focus of Machine Learning Applications for Health Research. *JAMA Netw Open*. 2019;2(10):e1914051. doi:10.1001/jamanetworkopen.2019.14051
20. Voller Sebastian, Mateen Bilal,A, Bohner Gergo, Király Franz,J, Ghani Rayid, Jonsson Pall et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness BMJ 2020; 368 :l6927
21. Linda Nevin, Advancing the beneficial use of machine learning in health care and medicine: Toward a community understanding, Published: November 30, 2018https://doi.org/10.1371/journal.pmed.1002708
22. Bhardwaj, R., Nambiar, A.R., & Dutta, D. (2017). A Study of Machine Learning in Healthcare. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 02, 236-241.
23. R. Bhardwaj, A. R. Nambiar and D. Dutta, "A Study of Machine Learning in Healthcare," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, 2017, pp. 236-241, doi: 10.1109/COMPSAC.2017.164.
24. K. Raghuraman, M. Senthurpandian, M. Shanmugasundaram, Bhargavi and V. Vaidehi, "Online Incremental Learning Algorithm for anomaly detection and prediction in health care," *2014 International Conference on Recent Trends in Information Technology*, Chennai, 2014, pp. 1-6, doi: 10.1109/ICRTIT.2014.6996092.
25. S. S. Khan, R. Barve and U. Kulkarni, "Proposed model on Prediction and Analysis using application of Health care," *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2018, pp. 340-345, doi: 10.1109/CESYS.2018.8723946.

26.Z. Yaowen, X. Wei and H. Yuwan, "Research on Healthcare Integrating Model of Medical Information System Based on Agent," *2010 International Conference on Computational and Information Sciences*, Chengdu, 2010, pp. 622-625, doi: 10.1109/ICCIS.2010.157.

27.N. A. Farooqui and R. Mehra, "Design of A Data Warehouse for Medical Information System Using Data Mining Techniques," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 2018, pp. 199-203, doi: 10.1109/PDGC.2018.8745864.