Predictors of Outpatients' No-Show: Big Data Analytics using Apache Spark

Tahani Daghistani (Daghistanita@ngha.med.sa)
Ministry of Naional Guard Health Affairs https://orcid.org/0000-0002-0268-2629
Huda AlGhamdi
Ministry of National Guard Health Affairs
Riyad Alshammari
King Saud bin Abdulaziz University for Health Sciences
Raed H. AlHazme
Ministry of National Guard Health Affairs

Research

Keywords: No-Show, Outpatient Clinics, Prediction Model, Big data, Spark

Posted Date: June 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-33216/v1

Version of Record: A version of this preprint was published on December 9th, 2020. See the published version at https://doi.org/10.1186/s40537-020-00384-9.

Abstract

Outpatients who fail to attend their appointments have a negative impact on the healthcare outcome. Thus, healthcare organizations facing new opportunities, one of them is to improve the quality of healthcare. The main challenges is predictive analysis using techniques capable of handle the huge data generated. We propose a big data framework for identifying subject outpatients' no-show via feature engineering and machine learning (MLlib) in Spark platform. The aim of this paper is exploring factors that affect no-sow rate then can be used to formulate predictions using big data machine learning techniques.

Introduction

Outpatient no-shows, who failed to attend scheduled appointments, in healthcare systems remain problematic [1, 2]. Patients' no-shows typically result in increased healthcare costs, underutilized medical resources and affect patient care [1, 3 4]. Clearly projected no-show increases is a key area for containing health care costs and improve system efficiency [3]. Healthcare organizations must consider the probability of patient no-show when scheduling appointments [3]. The performance of traditional strategic such as overbooking may not be consistently high, because it strongly depends on the no-show moments. In contrast, the use of machine learning to predict a no-show probability will guide decision for more reliable appointment scheduling strategies [5]. Predicting the patients who are most likely to miss their appointment can guide the facility towards better direction and care. It is worthwhile noting that this no-show case has been traditionally analyzed using historical data. The prediction techniques in the other fields such as economics already have a foundation, scientific findings and long history. However, these techniques are unusual within the healthcare, especially when restricted to the public domain [6].

A good starting point to achieve this goal is to explore factors that affect no-show rate based on information available, both for patients and appointments. Anticipated knowledge of patients' behavior is important, so that care clinics can react accordingly [6]. Since databases are large, they can exceed one million appointments for all Ministry of National Guard Health Affairs (MNGHA) facilities. The emergence of machine learning techniques along with big data analytics play a crucial role here has made it possible to carry out this study. Machine learning is an application of Artificial intelligence has been used widely by the research community to turn a variety, heterogeneous, huge data sources into high quality knowledge. Therefore, maximizing efficiency and discover cost-effective opportunities which consider as major pillar by healthcare providers [6, 7]. In addition to providing premier capabilities to discover pattern discovery or identify risk factors. However, applying machine learning techniques on complex big data is computationally expensive, it requires a massive computing power in terms of file space, memory, and CPU. A platform for big data analysis is becoming important as the data amount grows. Apache Spark MLlib is a platforms for big data analysis which offers a library for different machine learning techniques. In this contribution, we highlight big data machine learning from the computational perspective [7].

To handle increasing demand and recompense patient no-shows appointments, this paper provides a framework using big data to develop predictive models that identify the factors that influence a patient no-show. We explore the power of using Big Data Machine Learning models to accomplish this task.

Related Work

Several articles from other studies focusing on the various aspects of no-show in hospitals and documenting the effort to reduce no-show rate. Blumenthal et al. study aimed to develop a model to predict no-show for a scheduled colonoscopy. The predictive model used natural language processing (NLP) using historical medical records and endoscopy scheduling system. The model achieved AUC = 70.2 and 33% and 92% for sensitivity and specificity respectively [8]. Kurasawa et al. used logistic regression to predict missed appointments for diabetes patients. The value of AUC for the best predictor was 0.958; precision, recall and F-measure were, respectively, 0.757, 0.659 and 0.704 [9]. Devasahay et al. used historical appointment data merged with distance variable to predict no-show patients. They run Logistic Regression (LR), Support Vector Machine (SVM) and Recursive Partitioning to come up with predictive models. The best model was decision tree with 23.22% sensitivity and PPV of 15.58% (cut off.15). They were not be able to predict the type of patients will miss appointments accurately [10]. Goffman used logistic regression to model demographic and appointment characteristics, and the history of patient's behavior. The model accurately identified no-show patients with average AUC = 0.71 [11]. Harvey et al. used logistic regression to determine whether the patient successfully attend the appointment in the radiology department. The model considered 16 associated factors with AUC of 0.75. Further analysis was conducted based on different modalities; the predictive ability of the models were 0.74, 0.78 for CT and MAMMO respectively, and 0.75 for both MRI and ultrasound [12]. Elvira et al. proposed a new model that used Gradient Boosting (GB) algorithm for predicting no-show probability. A value of 0.74 for the Area under the curve (AUC) was the best results [6]. Srinivas and Ravindran proposed framework to develop no-show prediction models then proposed a scheduling rules using healthcare data from various sources. Among five different machine-learning algorithms used, stacking was the best with AUC = 0.846. Further, they integrated the no-show risk obtained from stacking model to the scheduling rules, this leads to improve the operational performance compared to the traditional overbooking approach [13]. Mohammadi et al. proposed three machine-learning models to predict no-show of next medical appointment. The overall accuracy of naïve Bayes was the highest, the model achieved 82%. The AUC for logistic regression, naïve Bayes and Multilayer perceptron are, 0.81, 0.86 and 0.66, respectively [14]. Dantas developed a predication model using logistic regression with an accuracy of 71%. The purpose of this model was to explore the factors related to no-show rates. They found that factors significantly associated with no-show in a bariatric surgery clinic were specialty, lead-time, the hour and month of the appointment, previous appointment and no-show history, type of appointment and distance [15]. Nelson et al, 2019 proposed predictive models for imaging appointments. They used four different algorithms, which are logistic regression, support vector machines, random forests, AdaBoost. The Gradient Boosting models achieved the best performance with AUC of 0.852 and precision of 0.511 [16]. AlMuhaideb et al. applied JRip and Hoeffding algorithms on historical outpatient scheduling data to build

predictive models. The predictive ability of both JRip and Hoeffding models were 76.44% and 77.13%, respectively, with area under the curve for JRip at 0.776 and for Hoeffding tree at 0.861 [17].

On the other hand, deep learning methods have attracted many researchers and organizations in health care field. Deep learning methods are useful with problems, which are difficult to solve with traditional methods. They are provide the optimal way to deal with high dimensional and volume data. Furthermore, present a whole picture embedded in large-scale data and disclose unknown structure. It has proven to be superior prediction of no-show thus effective optimizing of the health resource usage. There is very little effort in using deep learning in the prediction of patient's no-show. We have only found one study using deep learning to predict no-show patients in outpatients' clinics. Dashtban and Li, 2019 represented a novel prediction method for outpatients non-attendance based on wide range of health, environment and social economics factors. The model is based on deep neural networks, which have integrated data reconstruction and prediction steps from in-hospital data. This integration aiming to have higher performance than separated classification model in predicting tasks. The result of compare proposed model with other machine learning classifiers showed deep learning model outperforms other methods in practice. The model achieved (AUC (0.71), recall (0.78), accuracy (0.69). Finally, the constructed model was deployed and connected to a reminder system [18].

Method

Data for this study were extracted from Ministry of National Guard Health Affairs (MNGHA) data warehouse, a large institutional database derived from Electronic Medical Record (EMR). A total of (2,011,813, 19) data were queried for all outpatients visits scheduled from 2018 to July 2019 in the central region. All cancelled visits were excluded from the present study. The no-show factors have categorized in two groups. The first one involves appointments characteristics, as the appointment time, lead-time and distance. Second factors related to patients themselves, age, gender and the history of previous appointments. We also added calculated variables that will allow us to add information, such as the number of previous appointments, the number of no-show appointments, lead-time (number of days between reservation data and the appointment). The final group of attributes consisted of 20 attributes that were selected and calculated based on knowledge and previous work.

Data set are then pre-processed, eliminating incomplete and incorrect records, dealing with missing values and solving inconsistencies. Transformation between categorical or numerical data types was performed by means of normalization or scaling. In normalization, rescaling the attribute value from the original range to keep the values range between [0, 1]. In discretization, the age numerical attribute is transformed into a categorical attribute by selecting five as a cutoff point. Then we applied VectorAssembler function that transform all columns, both raw and calculated, into a single vector column can be passed to the ML algorithm [19]. Furthermore, we identified factors that have the greatest importance on the prediction and significantly influence the performance of the model. Information gain method has been used to rank features based on their impact on the show and no-show of patients and remove irrelevant features.

The collected data were divided into two sets in the ratio of 70:30, a training set including (1,408,269) for making model and a test set including remaining (603,544) as test set for validating and evaluating the model. As part of our work, we run an experimental evaluation of Apache Spark and MLlib under python programming languages using PySpark [20]. This study involves three machine learning techniques for predictive data task. That includes Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Support Vector Machine (SVM) and Multilayer Perceptron (MLP).

Random Forests (RF) is one of decision-tree classification algorithms that produce a different models (forest of trees) by selecting one input attribute randomly at each iteration and learning whether the classification results is more or less. At next iteration(s), the attribute either removed or included depending on the results of the previous iteration(s) [21]. Gradient Boosting (GB) method has been used in regression and classification. It is an ensemble of a number of weak decision trees prediction models to become a stronger learners. The prediction model resulted from GBM builds up in a stage-wise manner by adding new weak learners using a gradient descent to minimize the loss of the model. In boosting, a new learner is fit a subsample of the training dataset where selected randomly without replacement of full data set, then compute the model update for the current stage [22, 23]. Logistic regression (LR) is a one of predictive analysis methods that used to model the probability of a binary target. It use a linear combination of the different types of inputs and passes through the logistic function. Making predictions using logistic regression is easy to implement and provides a good results [24]. Support vector machine (SVM) algorithm is capable of constructing an optimal hyperplane to separate data points into classes based on a priori features of the training dataset. There are various hyperplanes or kernel functions that could be chosen in order to maximum distance between data points of both classes, so that future data points can be classified more accurately. The main advantages of SVM are the effectiveness in an N-dimensional space and it is memory efficient because it partition data into training points called support vectors used in the decision function [21, 25]. A Multilayer Perceptron (MLP) is an artificial network of neurons called Perceptrons. The perceptron computes a single output through nonlinear activation function from linear combination of multiple weighted inputs. Each Perceptron combined with many other perceptions and forms a fully connected network with input, output and hidden layers in between [26].

Matrices that used to select the best model are Accuracy, Precision, Recall, Fmeasure and Area Under the Curve, and F-measure. In addition to as well as the training and evaluation time, each metric is defined as follows:

- Accuracy: number of visits correctly classified.
- Precision: number of visits correctly classified by the system divided by number of all visits correctly classified by the system.

$$\text{Precision} = \frac{\text{TruePositive (TP)}}{\text{TruePositive (TP)} + \text{FalsePositive (FP)}} \qquad (1)$$

- Recall: number of visits correctly classified by the system divided by number of positive visits in the testing set.

$$\text{Recall} = \frac{\text{TruePositive (TP)}}{\text{TruePositive (TP)+ FalseNegative (FN)}} \qquad (2)$$

- F- measure: measure Recall and Precision at the same time, it represents the balance between both.

$$\text{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (3)$$

- ROC: measure classification performance at various thresholds settings by show how much model is capable of classify visits. It considers the tradeoffs in precision and recall [27].
- Time: training and evaluation time of the algorithms.

Results

From the (2,011,813) visits included in the study, there were (1,474,391) no-show (537,422) show visits. Therefore, the overall proportion of no-shows at all outpatients' clinics was (26.71%). Of these visits, we will not consider cancelled appointments. Each record contains 20 variables, which summarized in Table 1. As per Table 1, male patients were less likely to miss their appointments than female patients. New patients were the most likely to miss of their appointments. The patients who has Follow up were the second most likely to miss their appointments.

Table 1
Descriptive characteristics of the dataset (N = 2,011,813)

Features	No-Show (N%)	Show (N%)	Total (N)
Gender			
Male	213,729 (10.62%)	564,024 (28.04%)	777,753
Female	323,693 (16.09%)	910,367 (45.25%)	1,234,060
Age Group			
0–5	66,118 (3.29%)	150,748 (7.49%)	216,866
6–10	39,234 (1.95%)	95,566 (4.75%)	134,800
11–15	32,949 (1.64%)	87,269 (4.34%)	120,218
16–20	32,440 (1.61%)	87,115 (4.33%)	119,555
21–25	40,968 (2.04%)	104,714 (5.20%)	145,682
26–30	44,580 (2.22%)	118,409 (5.89%)	162,989
31–35	45,776 (2.28%)	123,426 (6.14%)	169,202
36–40	40,400 (2.01%)	114,836 (5.71%)	238,262
41–45	32,026 (1.59%)	95,960 (4.77%)	127,986
46–50	31,599 (1.57%)	97,445 (4.84%)	129,044
51–55	30,485 (1.52%)	94,975 (4.72%)	125,460
56–60	27,602 (1.37%)	88,763 (4.41%)	116,365
61–65	23,159 (1.15%)	74,015 (3.68%)	97,174
66–70	16,748 (0.83%)	51,621 (2.57%)	68,369
71–75	13,857 (0.69%)	39,771 (1.98%)	53,628
76–80	10,223 (0.51%)	27,134 (1.35%)	37,357
81–85	5,464 (0.27%)	13,272 (0.66%)	18,736
> 85	3,794 (0.19%)	9,352 (0.46%)	13,146
Nationality			
Saudi	530,112 (26.35%)	1,451,144 (72.13%)	1,981,256
Non-Saudi	6,650 (0.33%)	21,807 (1.08%)	28,457

Features	No-Show (N%)	Show (N%)	Total (N)
Unknown	660 (0.03%)	1,440 (0.07%)	2,100
Appointment type			
New Patient (NP)	243,158 (12.09%)	890,110 (44.24%)	1,133,268
First visit (FV)	271,466 (13.50%)	517,688 (25.73%)	789,154
Follow up (FU)	22,798 (1.13%)	66,593 (3.31%)	789,154
Reservation type			
Scheduled	516,300 (25.66%)	1,278,602 (63.55%)	1,794,902
Walk-in	21,122 (1.05%)	195,789 (9.73%)	216,911
Patient type			
Patient Service	530,923 (26.40%)	1,460,373 (72.59%)	1,991,296
Business Center	2,625 (0.13%)	6,634 (0.33%)	9,259
VIP	3,874 (0.19%)	7,384 (0.37%)	11,258
Distance (km)			
distance < = 100	517,591 (25.73%)	1,421,338 (70.65%)	1,938,929
distance > = 101 and distance < = 399	10,251 (0.51%)	28,962 (1.44%)	39,213
distance > = 400 and distance < = 799	7,462 (0.37%)	19,229 (0.96%)	26,691
distance >= 800	2,118 (0.11%)	4,862 (0.24%)	6,980
Outpatient Clinics			
Health Care Specialty Clinic	236,668 (11.76%)	656,152 (32.61%)	892,820
National Guard Comprehensive Specialized Clinic	102,428 (5.09%)	264,979 (13.17%)	367,407
King Abdulaziz City Housing	106,727 (5.31%)	304,584 (15.14%)	411,311
King Saud city Housing	81,487 (4.05%)	215,946 (10.73%)	297,433
Prince Bader Housing City Clinic	10,112 (0.50%)	32,730 (1.63%)	42,842

As an outcome of the feature importance process, the top four predictors are; number of no-show appointments, medical department, lead time and number of show appointments. The second four important predictors group are appointment type, patient type, outpatient clinics and appointment month. While appointment year, distance, gender, reservation type and nationality are not important predictors, thus removed from the models. The rest factors have less influence on the no-show such as number of schedule appointments, number of walk-in appointments, appointment time and age. The factors related to patients have more impact on no-show f patients to than factors related to the appointments. The prediction models have been developed using only 14 factors. The list of the predictors is ranked base on their importance in Fig. 2

Table 2 describes the experiments results carried out to show the performance of Spark using five machine learning algorithms over the same huge dataset. We evaluated the effectiveness of all classifiers in terms of time to train and evaluate the models, accuracy, precision, recall, F-measure and ROC. MLP and RF classified visits well. From the results, we can see that the percentage of all metrics is comparable for both classifier. A more improvement observed for the MLP in F-measure than RF. LG and SVM have similar ROC performance, LG are preferred than SVM as it produces better performance in all metrics with less computation power. SVM likely performs poorly due to the limitation of kernel function in MLlib, the only available linear kernel is used with SVM algorithm. GB performed best, resulting in an increase of accuracy and ROC to 79% and 81%, respectively.

Table 2
Evaluation metrics shown by different models on predicting outpatients no-show

	Accuracy	Precision	Recall	F-measure	ROC Area
Random Forest	0.76	0.76	0.76	0.68	0.77
Gradient Boosting	0.79	0.77	0.79	0.76	0.81
Logistic Regression	0.75	0.73	0.75	0.70	0.73
SVM	0.73	0.70	0.73	0.62	0.73
Multilayer Perceptron	0.77	0.75	0.77	0.72	0.78

To better understand efficiency, Fig. 2 presents the ROC curve of five models to illustrate the precision of each classifier. From the plot, we can easily show that Gradient Boosting is best model (area = 081). SVM with linear kernel and Logistic Regression returned comparable classification results. Currently, MLlib supports linear SVMs only; using non-linear kernels may outperform Logistic Regression.

As evaluation criteria, we have employed the overall training and evaluation time (in seconds) for all five algorithms. Table 3 compares time values obtained by five algorithms. Unlike other metrics, there are a differences between times of the algorithms and considered a huge difference in the training time. We observe that GB is around 15x times slower than MLP, although it achieved the optimal results. SVM, the algorithm with close performance to LG, takes about 68x times as long to train the model. Logistic

Regression is 4x times faster than the next two accurate algorithm MLP and RF with comparable performance. For huge datasets, the time is a factor to select one of the quicker algorithms, considering that the time values of models depends on the choice of algorithms parameters.

Table 3
Evaluation of time value for each machine learning model (seconds)

	Training Time	Evaluation Time
Random Forest	41.289	31.118
Gradient Boosting	668.882	27.287
Logistic Regression	10.033	24.962
SVM	685.782	23.081
MLP	42.444	23.458

Discussion

In this study, we attempt to identify the key factors to predict patients who will not attend the appointment (no shows) using regular available hospital data. The literature about predicting no-show has showed that logistic regression analysis was the main technique that used to identify factors influence no-show behavior. To the best of our knowledge, there are limited publications about the predication of no-show behavior using big data machine-learning approach such Spark. This study is designed to apply big data machine learning techniques in analyzing discrete patient EHR data to predict no-show probability based on Saudi Health Data.

There have been several studies focusing on the reason the patients' no-show. The main factor of no-show with the reasons such as mistakes and misunderstandings is the forgetfulness [28]. Other important factors for no-show were booking difficulty, work commitment, distance and seeking care in another healthcare facility [29]. Transportation is a key factor in addition to environmental factors that affect patient attendance and have value in predicting no-show. Factors including weather, distance, socioeconomic status and number of show in previous appointments [30]. Knowing factors associated with no-show can help improve quality of care and attempt to control factors that can be changed to reduce the no-show rate. This would have a direct impact on healthcare care in practical and financial way [29]. The results of this study recommend that reducing no-show rates among outpatients might be addressed by reviewing lead time specially it one of factors that controls by clinics.

Learning from previous studies, is clear that different interventions have a high success rate in reducing the negative impact of no-show [31]. A study conducted by Goffman indicated a reduction of no-show rate from 35–12.16%. The predicted no-show patients received a reminder call before 24, 48, and 72 hours of their appointments [11]. Arora et al. evaluated the effective of automated text message as

reminder system to increase show rate of follow-up appointment for patients after discharged from the emergency department. They found that the intervention was effective and reduced the overall appointment attendance rate from 72.6–70.2% [32]. Cancellation policies is one of intervention strategies to reduce patient no-shows and important for service operations. This could be used by clinics for rescheduling appointments. The findings indicate that when fill rates are low and no-show probabilities are high, the time required for patients to cancel appointments needs to increase in order to achieve the goal of being cost-effective [33]. A number of healthcare systems implemented SMS text messages as a reminder, which shows promise as an instant, simple, cost-effective means communications with protecting patient privacy. However, sending SMS to all patients, who have scheduled appointments, is not free. Using a prediction system will limit the sending of SMS to predicted show patients only. This would mean a cost reduction without affecting of attendance ratios [34, 35].

More studies are required investigating the extent the economic consequences of patient no-show and explored the factors that may modulate no-show rates. A real-world implementation of the model could validate our findings and assess the efficiency of the scheduling policy on patients' no-show behavior over time. There are various other factors can be explored and utilized for predicting no-show. Also, more improve seems to be plenty of room by attempting to add more features e.g. medication refill, lab appointments, or special clinic orders. Moreover, considered real-time predictive analytics is still an open question for future researches. Finally, The Spark cluster is setup using one node, further analysis is recommended by using multiple nodes.

Conclusion

Big data machine learning techniques have been shown to provide prediction capabilities in healthcare. In this study, we presented an evaluation of five machine learning techniques using Spark platform on predicting the patients' no-show using historical data from electronic medical records. Developing predictive models for the outpatients no-show could contribute to evaluating the relation between exposure to specific factors and the risk of no-show, to stratifying the patients' outpatient clinics with respect to this risk. This no-show risk model can be used to improve clinics' resource utilization and improve care access by recommending the overbooking of appointments to be occupied by patients at high risk of no-show.

Abbreviations

MNGHA: Ministry of National Guard Health Affairs; NLP: Natural Language Processing; LR: Logistic Regression; SVM: Support Vector Machine; GB: Gradient Boosting; AUC: Area under the curve; EMR: Electronic Medical Record; RF: Random Forest; MLP: Multilayer Perceptron; TP: True Positive; FP: False Positive; FN: False Negative.

Declarations

Availability of data and materials

Due to ethical restrictions imposed by the Institutional Review Board of King Abdullah International Medical research Center (KAIMRC), the data are available upon request to interested researchers.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the King Abdullah International Medical Research Center (KAIMRC), National Guard, Health Affairs, Riyadh, Saudi Arabia with research grant No. RC20/024/R.

Authors' contributions

TD developed the methodology, designed the experiment and took the lead in writing the manuscript. RS introduced the idea, contributed to data collection and preparation. HG and RH helped supervise and investigate the findings of this work. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Huang Y, Hanauer DA. Patient no-show predictive model development using multiple data sources for an effective overbooking approach. Applied clinical informatics. 2014;5(03):836–60.
2. Denney J, Coyne S, Rafiqi S. Machine Learning Predictions of No-Show Appointments in a Primary Care Setting. SMU Data Science Review. 2019;2(1):2.
3. Norris JB, Kumar C, Chand S, Moskowitz H, Shade SA, Willis DR. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. Decis Support Syst. 2014 Jan;1:57:428–43.
4. Samorani M, Harris S, Blount LG, Lu H, Santoro MA. Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. Available at SSRN 3467047. 2019 Oct 9.
5. Samorani M, LaGanga LR. Outpatient appointment scheduling given individual day-dependent no-show predictions. European Journal of Operational Research. 2015 Jan 1;240(1):245 – 57.
6. Elvira C, Ochoa A, Gonzalvez JC, Mochón F. Machine-learning-based no show prediction in outpatient visits. International Journal of Interactive Multimedia & Artificial Intelligence. 2018 Mar 1;4(7).

7. Assefi M, Behravesh E, Liu G, Tafti AP. Big data machine learning using apache spark MLlib. In2017 IEEE International Conference on Big Data (Big Data) 2017 Dec 11 (pp. 3492–3498). IEEE.
8. Blumenthal DM, Singal G, Mangla SS, Macklin EA, Chung DC. Predicting non-adherence with outpatient colonoscopy using a novel electronic tool that measures prior non-adherence. Journal of general internal medicine. 2015 Jun 1;30(6):724 – 31.
9. Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, Noguchi T, Ohe K. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. Journal of diabetes science technology. 2016 May;10(3):730–6.
10. Devasahay SR, Karpagam S, Ma NL. Predicting appointment misses in hospitals using data analytics. Mhealth. 2017;3.
11. Goffman RM, Harris SL, May JH, Milicevic AS, Monte RJ, Myaskovsky L, Rodriguez KL, Tjader YC, Vargas DL. Modeling patient no-show history and predicting future outpatient appointment behavior in the Veterans Health Administration. Military medicine. 2017 May 1;182(5–6):e1708-14.
12. Harvey HB, Liu C, Ai J, Jaworsky C, Guerrier CE, Flores E, Pianykh O. Predicting no-shows in radiology using regression modeling of data available in the electronic medical record. Journal of the American College of Radiology. 2017 Oct 1;14(10):1303-9.
13. Srinivas S, Ravindran AR. Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: a prescriptive analytics framework. Expert Systems with Applications. 2018 Jul 15;102:245–61.
14. Mohammadi I, Wu H, Turkcan A, Toscos T, Doebbeling BN. Data analytics and modeling for appointment no-show in community health centers. Journal of primary care community health. 2018 Nov;9:2150132718811692.
15. Dantas LF, Hamacher S, Oliveira FL, Barbosa SD, Viegas F. Predicting patient no-show behavior: a study in a bariatric clinic. Obesity surgery. 2019 Jan 15;29(1):40–7.
16. Nelson A, Herron D, Rees G, Nachev P. Predicting scheduled hospital attendance with artificial intelligence. NPJ digital medicine. 2019 Apr 12;2(1):1–7.
17. AlMuhaideb S, Alswailem O, Alsubaie N, Ferwana I, Alnajem A. Prediction of hospital no-show appointments through artificial intelligence algorithms. Ann Saudi Med. 2019 Dec;39(6):373–81.
18. Dashtban M, Li W. Deep learning for predicting non-attendance in hospital outpatient appointments.
19. Hung PD, Hanh TD, Diep VT. Breast cancer prediction using spark MLlib and ML packages. InProceedings of the 2018 5th International Conference on Bioinformatics Research and Applications 2018 Dec 27 (pp. 52–59).
20. Salloum S, Dautov R, Chen X, Peng PX, Huang JZ. Big data analytics on Apache Spark. International Journal of Data Science and Analytics. 2016 Nov 1;1(3–4):145–64.
21. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data
Dinov ID. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. Gigascience. 2016 Dec 1;5(1):s13742-016.

22. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002 Feb 28;38(4):367–78.
23. Mishra AK, Keserwani PK, Samaddar SG, Lamichaney HB, Mishra AK. A decision support system in healthcare prediction. InAdvanced Computational and Communication Paradigms 2018 (pp. 156–167). Springer, Singapore.
24. Jothi N, Husain W. Data mining in healthcare–a review. Procedia computer science. 2015 Jan 1;72:306 – 13.
25. Ahmad P, Qamar S, Rizvi SQ. Techniques of data mining in healthcare: a review. International Journal of Computer Applications. 2015 Jan 1;120(15).
26. Analysis of data mining techniques for healthcare decision support system using liver disorder dataset
Baitharu TR, Pani SK. Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. Procedia Computer Science. 2016 Jan 1;85:862–70.
27. Patel AC, Markey MK. Comparison of three-class classification performance metrics: a case study in breast cancer CAD. InMedical imaging 2005: Image perception, observer performance, and technology assessment 2005 Apr 6 (Vol. 5749, pp. 581–589). International Society for Optics and Photonics.
28. Neal RD, Hussain-Gambles M, Allgar VL, Lawlor DA, Dempsey O. Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. BMC family practice. 2005 Dec 1;6(1):47.
29. Alhamad Z. Reasons for missing appointments in general clinics of primary health care center in Riyadh Military Hospital, Saudi Arabia. International Journal of Medical Science and Public Health. 2013 Apr 1;2(2):258 – 68.
30. Mieloszyk RJ, Rosenbaum JI, Hall CS, Hippe DS, Gunn ML, Bhargava P. Environmental factors predictive of no-show visits in radiology: Observations of three million outpatient imaging visits over 16 years. Journal of the American College of Radiology. 2019 Apr 1;16(4):554-9.
31. Mohamed K, Mustafa A, Tahtamouni S, Taha E, Hassan R. A quality improvement project to reduce the 'No Show'rate in a paediatric neurology clinic. BMJ Open Quality. 2016 Aug 1;5(1):u209266-w3789.
32. Arora S, Burner E, Terp S, Nok Lam C, Nercisian A, Bhatt V, Menchine M. Improving attendance at post–emergency department follow-up via automated text message appointment reminders: A randomized controlled trial. Academic emergency medicine. 2015 Jan;22(1):31–7.
33. Huang Y, Zuniga P. Effective cancellation policy to reduce the negative impact of patient no-show. Journal of the Operational Research Society. 2014 May 1;65(5):605 – 15.
34. Foley J, O'Neill M. Use of mobile telephone short message service (SMS) as a reminder: the effect on patient attendance. European Archives of Paediatric Dentistry. 2009 Jan 1;10(1):15 – 8.
35. Parikh A, Gupta K, Wilson AC, Fields K, Cosgrove NM, Kostis JB. The effectiveness of outpatient appointment reminder systems in reducing no-show rates. The American journal of medicine. 2010

Jun 1;123(6):542-8.

Figures

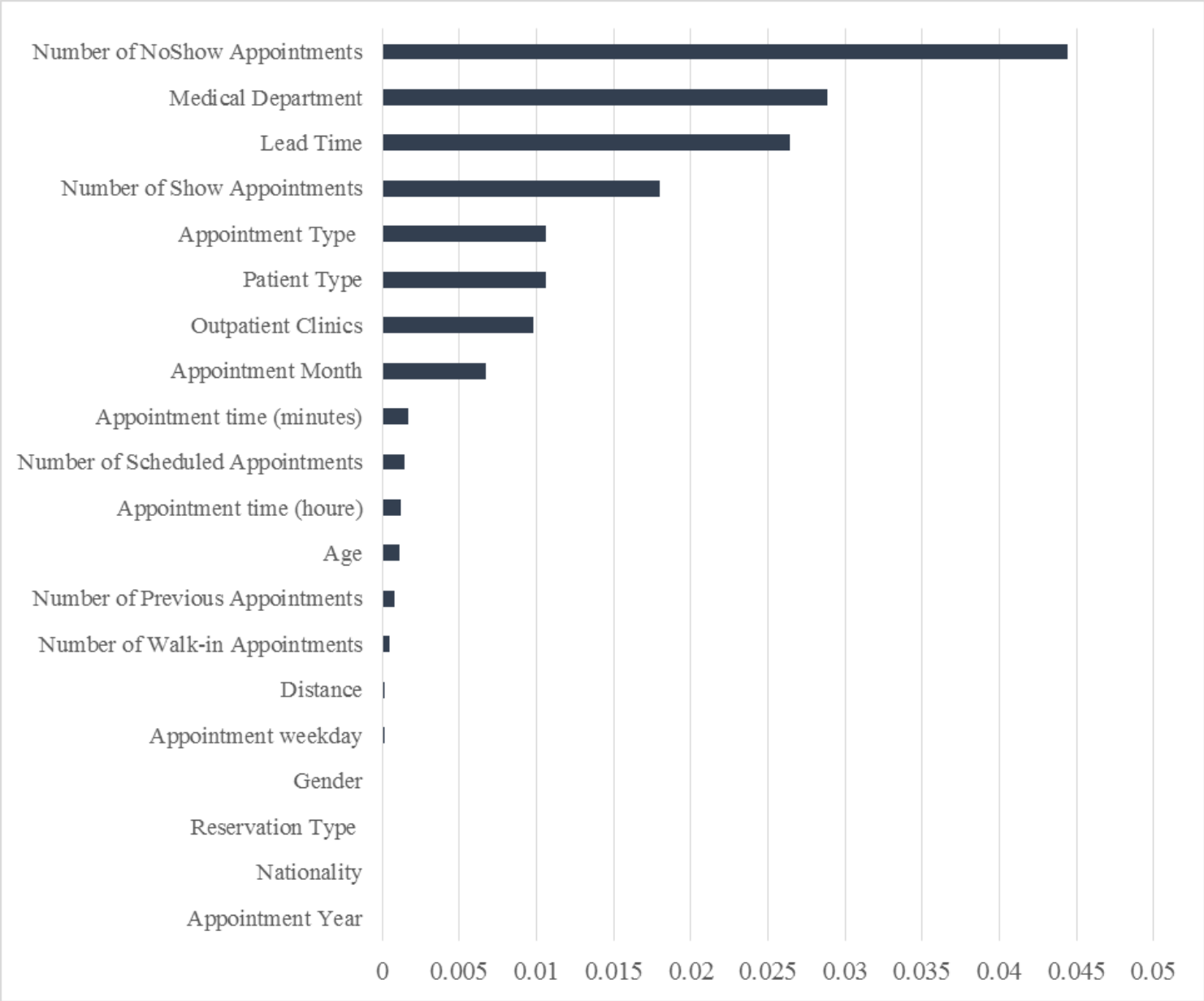

Figure 1

Feature importance ranking of factors in the developed machine learning models

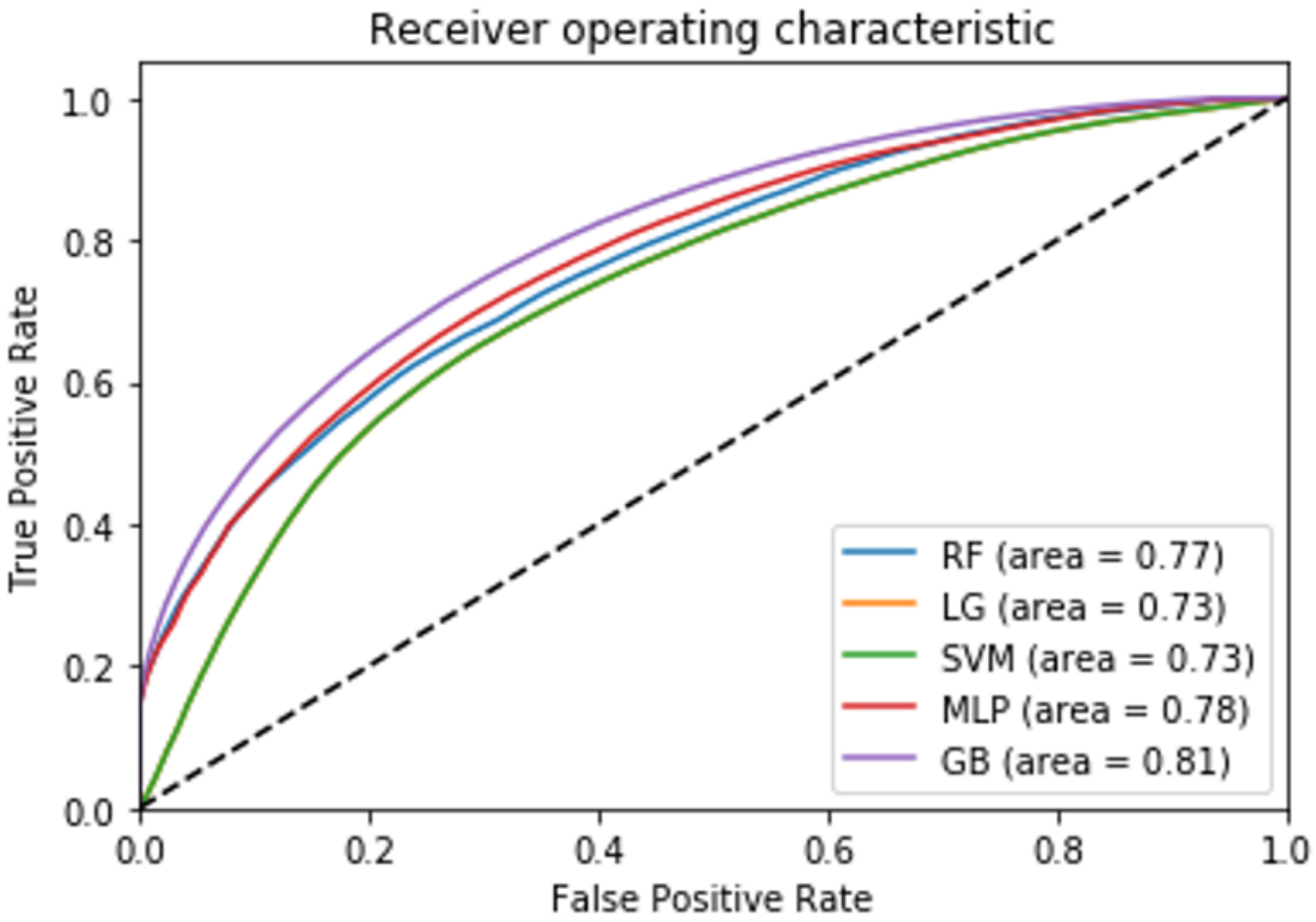

Figure 2

ROC of the developed machine learning models.