ISSN No:-2456-2165

Patient Attendance/No-Show Prediction

[1]Ooi Jien Leong
[1]MSc degree holder in Data Science & Business Analytics
Asia Pacific University of Technology & Innovation,
Kuala Lumpur, Malaysia

[2]Mafas Raheem
School of Computing
[2]Asia Pacific University of Technology & Innovation,
Kuala Lumpur, Malaysia

[3]Nowshath K Batcha
School of Computing
[3]Asia Pacific University of Technology & Innovation,
Kuala Lumpur, Malaysia

Abstract:- Patient no-show continues to contribute to the rising healthcare cost, leading to negative impacts on the day-to-day operations of the healthcare system, restricting healthcare delivery efficacy, besides limiting quality healthcare access for all patients. This study addresses the prevalence of patient no-shows, the missed by the patients. Demographic factors particularly age, gender, the time span of appointments and socio-economic status of patients are the most influencing factors on patient medical appointments attendance. Past attendance history, financial information, appointment information are among other factors that are also vital for patient attendance. Five machine learning predictive models namely Logistic Regression, Random Forest, Support Vector Machine, AdaBoost Classifier and Gradient Boosting Classifier were built using the 'Medical Appointment No-show' dataset after being treated for all possible types of noises. The Gradient Boosting Classifier was selected as the best performing model with 79.6% accuracy and 0.89 Receiver Operating Characteristics score as Gradient Boosting tends to perform better when it is properly tuned. Future research may include other key factors affecting patient attendance to improve model performance.

Keywords:- *Appointment Scheduling, Healthcare, Missed Appointments, Non-Attendance, Patient No-Shows, No-Shows Prediction, Predictive Models.*

I. INTRODUCTION

The global healthcare industry is a trillion-dollar industry. The ageing population, rapid growth of emerging economies, and changing social lifestyle are among the key factors contributing to the growth of the healthcare sector. Healthcare expenditure is projected to grow at 5.4% annually between 2017 and 2022, reaching approximately USD 10.06 trillion by the year 2022 [1]. Governments, employers and individuals are heavily burdened by rising healthcare costs, which continues to rise ahead of the inflation rate in many countries. In Malaysia, healthcare expenditure was RM 8,550 mil (USD 2,115 mil) in 1997 and increased to RM 22,072 mil (USD 5,500 mil) in 2006, and reached RM 57,361 mil (USD 14,200 mil) in 2017 [2]. Malaysia has experienced more than a five-fold increase in healthcare expenditure between 1997 and 2017. In this line, the missed appointment playing a significant role too. In the United States of America, patient non-attendance has created a loss of a whopping USD 150 billion annually to the country's healthcare sector [3]. Every missed appointment wasted 60 minutes of a medical officer at an average cost of USD 200 per appointment.

Patient no-show happens when a patient did not turn up to a prearranged medical appointment without cancelling or postponing the appointment with sufficient notice [4]. Patient no-show has been defined as patients who do not show up at the specified date, time, and location without giving notice [5]. Patients do not show up to medical appointments is prevalent. Diminished quality of healthcare system, increasing medical costs, inefficient manpower and resources allocation, deprived access to health services are among other negative consequences resulted from patient no-shows.

Ref. [6] concluded that high no-show is strongly associated with a significant economic cost. Patients who are absent from appointments and have to be treated with chronic care or emergency services at a later time, causing higher medical cost and straining the limited manpower and resources of hospitals. Subsequently, other major consequences of no-shows have decreased the productivity of the healthcare system, deprived access to equal and fair health services for other patients and disrupted disease management [7].

Despite many past studies on patient no-shows, a phenomenon of patients missing appointments persist and continues to cause the rising medical cost. Against this backdrop, this study aims to study the reduction of patient no-shows, hence suggesting improving quality of healthcare services and reducing medical cost. This research aims to conduct a comprehensive study in terms of both descriptive and predictive analytics and tries to achieve it via the following research objectives:

- To identify the underlying factors that influence patient attendance/no-shows to medical appointments.
- To build an effective predictive model that can predict patient no-shows to medical appointments.

- To draft valid recommendations for the relevant authorities in the healthcare industry to handle the situation more effectively.

Hence, effective handling of the no-show problem will bring practical benefits for the healthcare sector, and contributing to the world's economy as a whole when people have a fair opportunity to access healthcare services.

This paper is organised as Section 2 elaborates the literature reviews on patient no-shows. The methodology adopted in this study has been outlined in Section 3. Section 4 presents the results and discussions from this study. Section 5 concludes the study, followed by limitations and future works in Section 6.

II. LITERATURE REVIEW

Missed medical appointments have a significant impact on revenue and operating cost, affecting resource allocation of the healthcare system [8]. Missed medical appointments are positively correlated with multimorbidity and mortality [9] and the rates vary by different factors [10]. Ref. [11] reviewed 105 studies and discovered no-show rates at an average of 23%, where the rate varies by specialties and geographical continents.

A. Key Factors Influencing Patient No-shows

Several studies proved that demographic and medical factors are the key drivers to patient no-shows. Age, gender and socio-economic status are important factors influencing no-shows as well [12]. Patients who are females, from younger age group and poor socio-economic status are significant for causing non-attendance [13]. Travel distance, transportation, work commitment, and closer alternative healthcare facilities are some other key reasons for no-shows [13], [14]. Younger or older age group, male, having greater deprivation, suffering from suspected cancer site, referred to treatment at the early stages reported patient no-shows due to long distance to the medical centres. The race is also one of the demographic factors influencing patient attendance at medical appointments [15].

Ref. [16] concurred that males are more likely to miss medical appointments than females regardless of age groups. Further, patients of a certain profile: male, younger adults, low socio-economic level, no private insurance and long-distance from the clinic are more likely to be associated with a higher probability of no-show [11]. On the other hand, distance is not of major concern in a relatively small city-state with a good public transportation system for no-shows of medical appointments [17]. The patient groups such as patients with specific diseases and those who are undergoing several chronic diseases and the patients being treated with multiple chronic diseases do not constitute higher no-show probabilities [18]. Patients who are male, young, and from the low socioeconomic status as well as the unmarried and receiving welfare payment command a higher chance of missing medical appointments [7], [18].

Accordingly, the lead-time/time-span is a period between the appointment date and the actual appointment day, which is one of the common predictors of patient attendance followed by the demographic profiles. Lead time shows a positive correlation to patient no-shows to medical appointments [7], [11], [19]. Having multiple appointments on the same day is positively correlated with attendance rates too [7]. Information about appointment including time, day, and duration is also among the factors influencing attendance at medical appointments [15]. Internal referrals tend to have higher no-shows in general, as well as the appointment scheduled in the morning between 8 am to 9 am and afternoon between 12 pm to 2 pm tend to have lower no-show rates [16].

Class categories/financial information on whether or not a patient is subsidized or self-paying for medical treatment is one of the influencers to patient attendance [11], [16]. A subsidized patient showed higher no-shows than patients with private health insurance. Further, past attendance behaviour is an important factor influencing no-shows [7], [11], [15]. Human factors such as forgetfulness and communication errors on appointment time, appointment cancellations, fear of the medical procedure are some of the key factors influencing patient no shows [20], [21].

B. Patient Attendance/No-Show Prediction

There have been numerous studies that adopted statistical analysis to predict patient no-shows for medical appointments. However, only a handful of studies were investigated with the application of big data technologies to predict patient no-shows for medical appointments.

Logistic regression algorithm appears to be the most commonly used algorithm in the area of patient attendance prediction [7], [15], [16], [17], [22], [23]. Ref. [7] worked on building 24 unique predictive models using logistic regression and tested by making reminder calls to patients before an appointment. Their model accurately predicted patients with high no-show risk; targeted intervention strategy has successfully improved no show rate from 35% to 12.16% [7].

Ref. [22] trained a logistic regression model with a backward elimination criterion of $p<0.05$ to predict excessive missed medical appointments with multiple sclerosis. The predictive model was accompanied by a score indicating patients who tend to miss more than 20% of appointments. The model was useful for healthcare institutions to implement and triage intervention strategies to reduce non-attendance, besides, to encourage adherence to multiple-sclerosis related treatments.

A regularized logistic regression model was built and model performance was evaluated using the discrimination and calibration results [15]. Ref. [23] modelled patient no-show prediction using stepwise naïve and mixed-effect logistic regression and the best fit model generated good performance and passed the necessary validation procedure. This finding proved its effectiveness and usefulness, enabling healthcare staff to make an informed decision by overbooking in the appointment schedule.

Gradient Boosting algorithm was modelled by including only variables that are significantly associated with target appointments [24]. The model was trained further by considering appointment history into the modelling process.

Six predictive models namely decision tree, logistic regression, gradient boosting trees, random forest, elastic-net and XGBoost were built using feature engineering with text mining approach and evaluated based on 5-fold cross-validation [17]. XGBoost yielded the highest Area Under Curve (AUC) & precision score and chosen as the best predictive model according to their study.

Ref. [16] attempted to model appointment misses by applying logistic regression, decision trees and support vector machine (SVM) techniques. The decision tree was the selected algorithm for the prediction of appointment misses as it was the best fit model with a 0.15 cut off compared to logistic regression and SVM models [16].

Further, a numerical methodology was designed to evaluate outpatient schedules, by taking into consideration patients waiting time, idle time and overtime of doctors [25]. The evaluation method has been included in a local search algorithm, providing insights to admin staff when scheduling appointments for unpunctual patients. A good appointment system would be able to predict the unexpected situation that could arise and manage the disruption in the appointment scheduling process [8].

By utilizing social and demographic profile, together with appointments attendance records of patients, a hybrid probabilistic model based on logistic regression and empirical Bayesian inference was built to predict the real-time patient no-show. The model employed a precise selective overbooking strategy to reduce the impact of non-attendance. It was able to slots in appointment sessions while keeping a short waiting time. The authors later enhanced their probabilistic model, which was not only be used to estimate probabilities of no-show but also able to estimate cancellation and attendance in real-time [26].

Integrating elastic net variable-selection methodology into the probabilistic Bayesian Belief Network (BBN), a hybrid probabilistic prediction system was developed by extracting the relationship between no-shows' predictors and the conditional no-shows' probability [27].

Table 1 summarizes the patient attendance predictive models that were built by researchers along with respective model results.

TABLE I. PATIENT ATTENDANCE PREDICTIVE MODELS

No.	Author	Algorithm Used	Model Results
1.	Goffman et al., 2017 [7]	Logistic Regression	No-shows reduced from 35% to 12.16%
2.	Ding et al., 2018 [15]	Regularized Logistic Regression	• Proven the value of fitting local level models • Highlighted the importance of developing "personalized" risk scores
3.	Devasahay, Karpagam and Ma, 2017 [16]	Logistic Regression, Decision Trees, Support Vector Machine	Decision tree was the best fit model
4	Lee et al., 2017 [17]	Decision Tree, Logistic Regression. Gradient Boosting Tree, Random Forest, Elastic-Net, XGBoost	XGBoost was selected, with AUC of 0.832, Precision of 0.785
5	Gromisch et al., 2020 [22]	Logistic Regression - backward elimination	Model was able predict patients who will miss more than 20% of appointments
6	Lenzi, Ben and Stein, 2019 [23]	Stepwise Naïve, Mixed-effect Logistic Regression	The best model developed has AUC 80.9%
7	Elvira, Ochoa, Gonzalvez and Mochon, 2018 [24]	Gradient Boosting, General Linear Model, Deep Learning	Gradient Boosting algorithm with 74% AUC was selected
8.	Alaeddini, Yang, Reeves and Reddy, 2015 [26]	Hybrid probabilistic model based on logistic regression and empirical Bayesian	Built a model which predicts the real pattern correctly with small variance.

C. Intervention to Reduce No-show Rates

Patients' no-shows is a perennial issue for the healthcare system. Not only does it affect the quality of clinical care, but it also leads to inefficient use of resources, causing cost and time wastage when there is a delay and overcrowded situation in the waiting room or idle time of doctors.

Intervention strategies to reduce no-show can be broadly group into 3 main categories, namely reminders, reducing perceived barriers and increase motivation [20]. Reminders are a possible intervention to encourage patient attendance [11]. Given that forgetfulness is one of the main reasons for non-attendance where SMS reminders can boost up the show-up rates for patients who tend to forget about their appointments [28] [29]. Sending reminders via phone or text messages is effective in no-show reduction. However, setting up a reminder system requires the healthcare providers to have an accurate record of patient's contact numbers and have a telephone system which is capable of handling the extra calls [20].

Having flexible and an open-access appointment booking system for patients to make the booking and make changes according to their schedule is an effective way to reduce no-shows [11], [20], [30]. Given the outpatient offering and the fixed number of appointment slots, it is impossible to increase appointment slots. Therefore, healthcare intuitions should minimize the negative impact of a no-show by implementing appointment scheduling with overbooking [11], [19].

Better communication is important to ensure the improvement of healthcare access for urgent patients while reducing overall waiting time [30]. Leveraging on the business intelligence system, a business intelligence dashboard for intervention was proposed, which can generate a no-show risk score, enabling the implementation of risk-based no-show management in the clinics [17].

III. METHODLOGY

The ultimate goal of healthcare institutions is to deliver high-quality healthcare services to patients by delivering valuable patient care at the lowest cost. Medical appointments are sometimes missed by the patients due to various factors. The healthcare institutions try their level best to reduce the patient no shows records by employing tremendous human efforts. However, handling the no show issues manually is not feasible for long term operations. Therefore, employing data mining techniques seems very useful in handling these kinds of situation. A machine learning predictive model which can classify the outcome of patient no-shows as accurately as possible would be an ideal solution.

This study uses a real-world Medical Appointment No-show dataset acquired from 'Kaggle.com'. It comprises 110,527 appointment records collected from several medical centres in Vitória, Brazil in the year 2016 [31]. It consists totally of 14 variables, of which 13 are explanatory; and 1 target binary variable. Summary of the dataset is presented in Table 2.

TABLE 2. MEDICAL APPOINTMENT NO-SHOWS DATASET SUMMARY

Attribute Name	Description
PatientId	Identification of a patient
AppointmentID	Identification of each appointment
Gender	Male or Female
ScheduledDay	The day registered the appointment
AppointmentDay	The day of the actual appointment
Age	Age of the patient
Neighbourhood	Where the appointment takes place
Scholarship	True or False
Hypertension	True or False
Diabetes	True or False
Alcoholism	True or False
Handicap	True or False
SMS_received	True or False
No-show	True or False (Target Variable)

A. Data Preparation

Data preprocessing is an essential and time-consuming task in any data analytics project. A set of data preprocessing activities were comprehensively carried out such as data transformation, data type conversion, categorical data encoding and a new variable creation using Python to improve the data quality for the modelling. The details of the data preparation were discussed in Section IV.

B. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) helps to gauge the influence of each independent variables in predicting the target variable. The cleaned dataset was used to get insights from the data. EDA using charts and graphs were created to show the distribution of the dataset, enabling a better understanding of the relationship between probabilities of no-show against other variables.

C. Data Mining Techniques

Data mining is a powerful tool in the real world for business and non-business applications. It is the process of employing machine learning algorithms to uncover patterns in large datasets [32]. In the context of the healthcare domain, predictive models for the patient no-show is useful for healthcare operators to improve resource utilization, enabling healthcare systems to allocate precious clinic appointments to patients in a timely and efficient manner. Intervention strategies can be implemented based on the outcome of the predictive models to reduce patient no-show rates. The data mining goal of this study is to classify patient attendance/no-attendance to medical appointments.

Five classifiers namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), AdaBoost (AB) and Gradient Boosting (GB) were built with suitable hyper-parameter. The selection of the algorithms was made based on the literature and profitable existing hyper-parameters. Also, the most effective hyper-parameters were selected using GridSearchCV method. And the best performing predictive model was selected based on the model evaluation measures such as accuracy, precision, recall, F1 Score and ROC score as the final model for patient attendance prediction. The detailed modelling tasks were discussed in Section IV.

1) *Logistic Regression (LR):* LR is one of the primitive algorithms that utilize a sigmoid function and powerful on binary classification tasks. It is the most commonly used machine learning algorithm in many studies in the past. Hence, the LR model was developed in the study as the benchmark model for comparison since many researchers built LR models as per the literature review.

2) *Random Forest (RF):* RF is an ensemble machine learning technique that gives competitive accuracy on most datasets. It is robust, easy to use and proven effective at preventing overfitting issues. However, one of the disadvantages of RF models is that it may suffer from overfitting and not easily detected. Other performance evaluation metrics such as Receiver Operating Characteristic (ROC) curve and confusion matrix is better

to be used as an additional evaluation metrics along with the model accuracy.

3) *Support Vector Machine (SVM):* SVM algorithm is suitable for building classifiers in the absence of noise in a dataset, which does not contain overlapping target classes. Besides an advantage of memory efficient, SVM is capable of capturing complex relationships between data points without the need for complex variable transformation. Hence, data transformation would always support the algorithm for effective learning.

4) *AdaBoost Classifier (AB):* AB is an ensemble classifier that iteratively trains a model by selecting the training set based on the accuracy of previous training. AdaBoost is adaptive, where it corrects the previous errors by tuning the weights for every incorrect observation in every iteration. AdaBoost is suitable even for imbalanced datasets but underperforms when noise is found in the dataset. However, the algorithm was selected to build a model while the dataset was balanced and it found no noise.

5) *Gradient Boosting Classifier (GB):* GB is another famous ensemble classifier and also known as an additive model in a forward fashion. The model is built by allowing for the optimization of arbitrary differentiable loss functions. GB is more useful for binary classification where only a single regression tree is induced to build the model.

D. Research Framework

The research framework for building a predictive model of patient attendance is illustrated in Fig. 1 based on the findings from the literature reviews, and the exploratory data analysis.

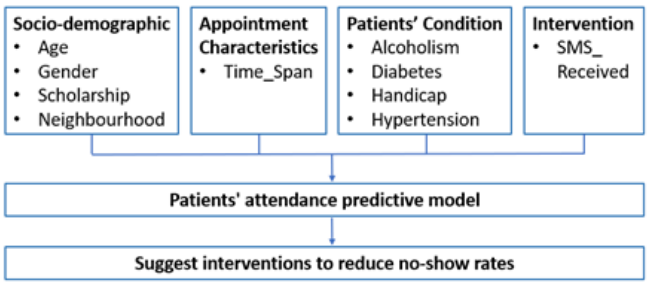

Fig. 1. Research Framework

The input variables were cleaned and set to build predictive models to predict patient attendance/no-attendance to medical appointments. Though few relevant past studies were found in this area, certain crucial preprocessing activities were taken to clean the data.

IV. RESULTS & DISCUSSION

A. Data Preparation and Pre-processing

Comprehensive data preparation was done where the data was checked for all types of noises. It was cleaned, recoded, transformed, and computed new variables in preparation for better use of dataset in the modelling stage. The cleaning of the dataset included renaming the incorrect variable names, dropping unwanted variables and dropping correlated insufficient records.

B. Feature Engineering

Feature engineering is the extraction of new features from raw data and transforming them into a new format that can better represent data. This is an important step in the data mining process to improve the quality of the machine learning model results [33].

A new variable, Time_Span was created from the variables AppointmentDay and ScheduledDay where it refers to the number of days a patient needs to wait for the actual appointment date after an appointment is scheduled. Also, the records showed negative values of Time_Span were being converted into positive values. Subsequently, the AppointmentDay and ScheduledDay were dropped along with the PatientId and AppointmentID from the dataset as these variables found to be no longer useful for the model building.

C. Feature Selection

Feature selection is the process of selecting the useful/impactful input variables for modelling to improve model performance. At this phase, a Pearson correlation heatmap was plotted to check the correlation among the input variables and with the target variable No_Show. The variables that obtained a correlation value above 0.9 were then been removed. However, no high correlated input variables were found in the dataset. So, the cleaned dataset was finalised with 11 variables in total.

D. Exploratory Data Analysis (EDA)

Interactive dashboards are powerful for business decision-makers to understand insights effectively. An EDA was performed to graphically examine the relationship of each variable with the target variable, No_Show.

1) *No_Show Patient Profile:* One in every five medical appointments were missed according to Fig. 2.

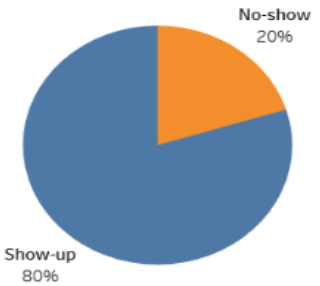

Fig. 2. Prevalence of No-shows

The dashboard in Fig. 3 depicts no-shows by socio-demographic factors such as Age Group, Gender and Scholarship. It shows that patients in the younger age group are more likely to miss their medical appointments. No-shows rate was declined with patients' age, and no-shows rate is lower among patients who are above 60 years old. Higher no-show rates are observed among patients who are

Scholarship receivers. However, no difference in no-show rates between male and female patients.

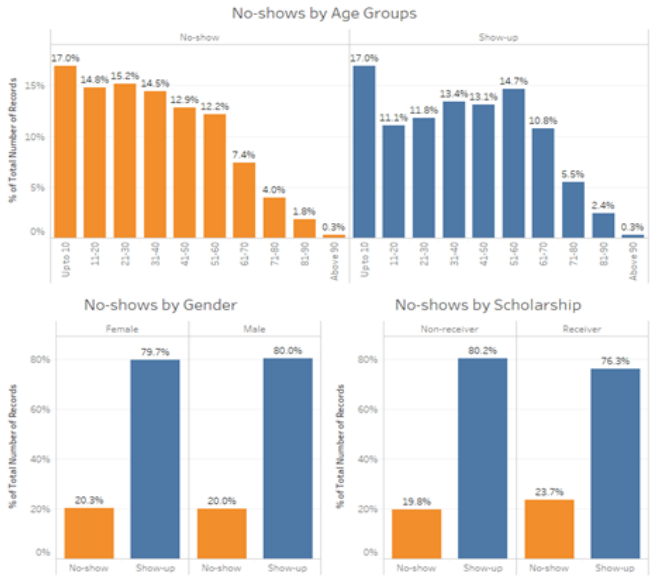

Fig. 3. No-Shows by Socio-demographics Factors

Fig. 4 is a dashboard that displays disease conditions of Hypertension, Diabetes, Handicap; as well as Alcoholism. It reveals that patients suffering from hypertension tend to have lower no-shows.

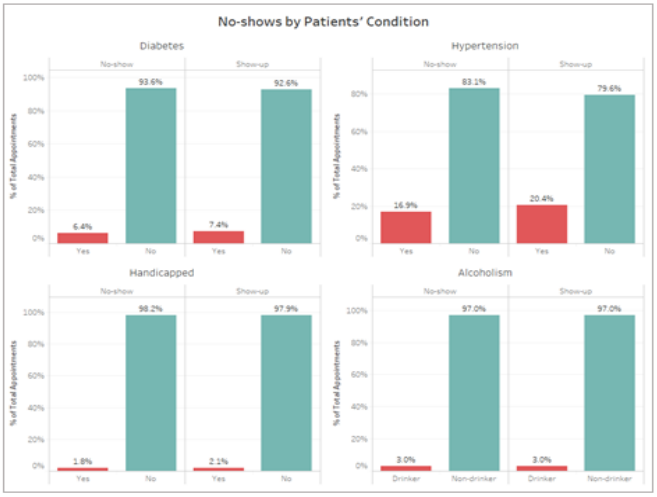

Fig. 4. No-shows by Patient Condition

Similarly, diabetic patients command a slightly lower no-shows rate compared to those without diabetes. Healthy patients are more likely to miss their appointments; while lower no-shows are observed among patients suffering from any one of these three disease conditions.

Timespan or lead time is positively correlated to no-shows, where longer lead time tends to cause a higher no-show rate. As illustrated in Fig. 5, lower no-shows for appointments that took place within the same day or on the next day from the day it was scheduled. Appointments with 2-7 days lead time, on the other hand, have the highest no-shows.

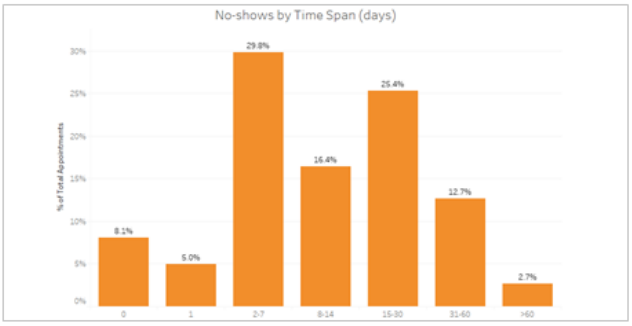

Fig. 5. No_Show by Time Span (in days)

2) Key Factors Influencing No_Show: Age and socio-economy appear to be significant predictors of patient attendance. However, gender is insignificant for attendance prediction. Lead time is one of the important features of influencing no-shows. The total number of diseases appears to be a significant predictor; patients suffering from diabetes and hypertension tend to command a lower no-show rate compared to those that are handicapped. The target intervention included in this dataset, i.e. SMS reminder seems not significant in affecting patient attendance. As shown in Fig. 6, higher no-show rates (43.8%) were observed compared to 29% who show-up for appointments among patients who have received SMS reminders.

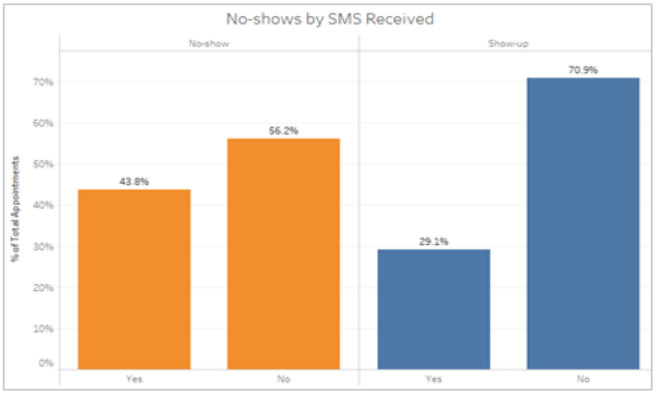

Fig. 6. No-shows by SMS Received

E. Handling Imbalance Class

Class imbalance is common in real-world, prevalent in supervised machine learning problems such as disease diagnostics, identification of fraudulent transactions, and detection of oil spills in satellite radar images [34]. The Medical Appointment No_Show dataset is imbalanced, with 20% instances of 'No_Show' in its target class. Classifiers with imbalanced data are foreseen to pose a challenge during modelling, providing misleading yet biased results by predicting only the majority class, i.e. 'No_Show' as the 'correct' way in almost every prediction.

Near Miss undersampling method was used to handle the imbalanced dataset distribution. This is an undersampling method that select instances based on the distance of majority class instances to minority class instances [35]. NearMiss-1 technique was used to select instances from the majority class ('show-up') that have the smallest average distance to the

three closest instances from the minority class ('no-show') as shown in Fig. 7.

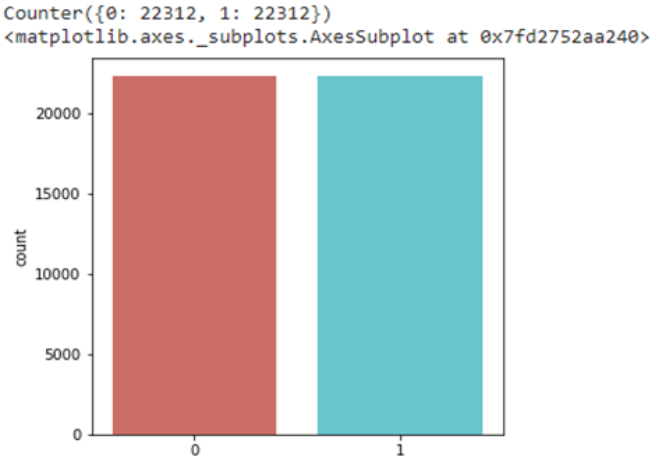

Fig. 7. Class balancing using under sampling

F. Data Normalization

The dominant features and outliers are the two possible issues found with the raw data that could hinder the learning process of machine learning algorithms. The data normalization is a well-known operation to either transform or rescales the raw data to get uniform contributions for the model building process. The data set was found with few outliers and normalized to get uniform contributions to build the selected machine learning models.

G. Data Modelling

Five classifiers namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), AdaBoost (AB) and Gradient Boosting (GB) were developed to build a predictive model with better accuracy. The data was split into 80% for training, and 20% for testing to provide an honest assessment of all the predictive models. The predictive models were built with the support of the GridSearchCV to select the most profitable hyper-parameter combinations. The hyper-parameters were selected using GridSearchCV with cross-validation using the test data, thus the final accuracy score can be reliable. Table 3 showcases the selected hyper-parameters of the respective models. The models obtained different accuracy scores based on the tuning of the hyper-parameters.

TABLE 3. HYPER-PARAMETERS OF THE MODELS

Model	Hyper-Parameters	Type
Logistic Regression	multi_class = 'ovr' class_weight: 'dict' random_state: 44 solver: 'saga' tol: 0.001	Traditional Machine Learning Algorithms
Support Vector Machine	C: 9 class_weight: 'balanced' gamma: 'scale' kernel: 'rbf'	
Random Forest	n_estimators: 900 min_samples_split: 10 min_samples_leaf: 4 max_features: 'auto' max_depth: 30 criterion: 'entropy' class_weight: 'balanced' bootstrap: True	Ensemble Machine Learning Algorithms
AdaBoost Classifier with Decision Tree as the base estimator	**Decision Tree** criterion = 'entropy' max_depth = 50 max_features = 'log2' min_samples_leaf = 1 min_samples_split = 100 random_state = 8 **Adaboost** learning_rate: 0.0001 n_estimators: 20 random_state: 2	
Gradient Boosting	loss = 'deviance' n_estimators = 1000 max_features = 'auto' criterion = 'friedman_mse' learning_rate = 0.1	

ISSN No:-2456-2165

Model performance metrics including accuracy, precision, recall, F1 score and ROC score for each of the five models were evaluated to select the best model for this purpose. However, as the balanced data was used to train and test the model along with cross-validation, the accuracy values of the respective models were plotted as shown in Fig. 8. According to the accuracy values, the RF, AB and GB models obtained good values. However, GB is the best performing algorithm selected in this study as it commands a slightly higher performance (Accuracy 79.6%) compared to RF (78.7%) and AB (78.6%).

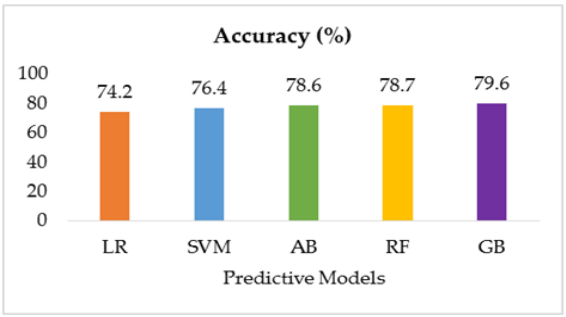

Fig. 8. Predictive model accuracy values

Based on the evaluation among the models, GB was selected as the most suitable model which obtained Receiver Operating Characteristics (ROC) score of 0.89 along with the respective confusion matrix (Fig. 9).

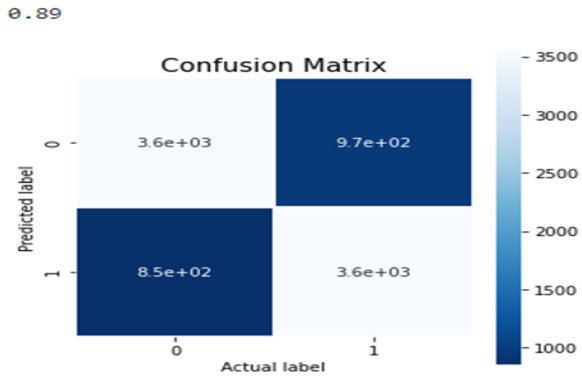

Fig. 9. ROC Score & Confusion Matrix of GB

Fig. 10 depicts the Receiver Operating Characteristics (ROC) curve with the index of 0.89 of the GB model.

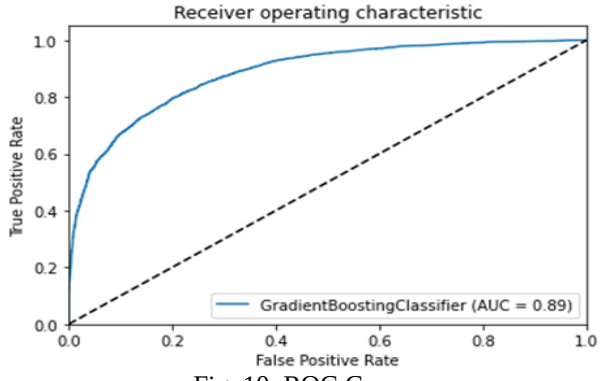

Fig. 10: ROC Curve

V. CONCLUSION

Patient No-shows for medical appointments is a prevalent challenge especially for all types of healthcare institutions. The challenge persists and continue to cause the escalating healthcare costs. This study aims to reduce patient no-shows, thus improving quality of healthcare services whilst reducing medical cost. It contributes to establishing a better understanding to the domain of patient no-shows. Descriptive predictive analytics provided new insights that are useful to suggest in reducing the patient no-shows.

An effective communication strategies can be developed with the insights gained from this study as the SMS reminder seemed not effective to reduce patient no-shows. As forgetfulness is one of the main reasons for non-attendance, it could be recommended to the healthcare institutions to implement 2-ways interactive reminder systems such as phone calls to remind and reconfirm with patients on the appointment date and time ahead of the appointment days. Leveraging the emerging technology such as emails and/or WhatsApp messages, the healthcare institutions can consider informing patients on the nature of the upcoming medical appointment, treatment plan, reasons for the required treatment. Given the potentially high cost of intervention implementation, healthcare institutions may roll out interventions only among the patients who are possibly absent for medical appointments.

Five machine learning predictive models namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), AdaBoost (AB) and Gradient Boosting (GB) were constructed. Models were assessed on their ability in achieving the defined data mining goal of classifying the patient no-shows with higher accuracy. GB was the selected best performing model with 79.6% accuracy and with 0.89 ROC score while confirming that this model outperforms all the models found from the literature review. GB used to offer a better performance when the algorithm is trained with noise free data and the model is properly tuned with profitable hyper-parameters.

As the application of classification/predictive models is gaining strength in the healthcare sector, further research can be done to determine the best performing model when additional impactful variables are included. A good predictive model will not only improve the effectiveness of the operations of healthcare institutions but also able to reduce the number of missed appointments that lead to high healthcare operational costs to the community. This will ultimately aid to achieve the goals of healthcare institutions to deliver high-quality healthcare services to all patients, through delivering high-value patient care at the lowest cost possible.

VI. LIMITATIONS & FUTURE WORK

This study limits only to features contained in this dataset thus are lacking some significant variables such as human behaviour, distance from the house, transportation mode which may be included in the future study. Furthermore, the target variable in this dataset is a binary class of show-up and no-shows where the cancelled appointments were not captured due to any valid reasons. It is, therefore, suggested for future study to also include records of cancelled appointments. Subsequently, it would be useful to identify patients who are likely to cancel their appointment last minute, as well as patients who are usually late to appointments to ensure effective resources allocation well in advance. In this context, future works can also consider including the time of appointment cancellation and lateness duration for better understanding of patient no-shows. Also, the Deep Learning architectures may be deployed in building predictive models with higher accuracy for future use.

Establishing an understanding of no-shows and the number of missed appointments prediction would enable healthcare institution operators to manage appointment scheduling system effectively.

REFERENCES

[1]. Allen, S. (2019). 2019 Global health care outlook | Shaping the future. [Online] 2019 Global healthcare outlook. Available from: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-hc-outlook-2019.pdf [Accessed: 15 Dec. 2019].

[2]. Ministry of Health Malaysia (2019). Malaysia National Health Accounts: Health Expenditure Report (1997-2017). 1st ed. Putrajaya: Ministry of Health Malaysia.

[3]. Gier, J. (2019). Missed appointments cost the U.S. healthcare system $150B each year. [Online] Healthcare innovation. Available from: https://www.hcinnovationgroup.com/clinical-it/article/13008175/missed-appointments-cost-the-us-healthcare-system-150b-each-year [Accessed: 10 Dec. 2019].

[4]. Lacy, N. (2004). Why We Don't Come: Patient Perceptions on No-Shows. The Annals of Family Medicine, 2(6), pp.541-545.

[5]. Blæhr, E., Kristensen, T., Væggemose, U. and Søgaard, R. (2016). The effect of fines on nonattendance in public hospital outpatient clinics: study protocol for a randomized controlled trial. Trials. 17(1).

[6]. Kheirkhah, P., Feng, Q., Travis, L., Tavakoli-Tabasi, S. and Sharafkhaneh, A. (2016). Prevalence, predictors and economic consequences of no-shows. BMC Health Services Research. 16(13).

[7]. Goffman, R., Harris, S., May, J., Milicevic, A., Monte, R., Myaskovsky, L., Rodriguez, K., Tjader, Y. and Vargas, D. (2017). Modeling patient no-show history and predicting future outpatient appointment behavior in the Veterans Health Administration. Military Medicine. 182(5). pp.e1708-e1714.

[8]. Alaeddini, A., Yang, K., Reddy, C. and Yu, S. (2011). A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Management Science. 14(2). pp.146-157.

[9]. McQueenie, R., Ellis, D., McConnachie, A., Wilson, P. and Williamson, A. (2019). Morbidity, mortality and missed appointments in healthcare: a national retrospective data linkage study. BMC Medicine. 17(2).

[10]. Salemi Parizi, M. and Ghate, A. (2016). Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. Computers & Operations Research. 67. pp.90-101.

[11]. Dantas, L., Fleck, J., Cyrino Oliveira, F. and Hamacher, S. (2018). No-shows in appointment scheduling – a systematic literature review. Health Policy. 122(4). pp.412-421.

[12]. Pepino, A., Vallefuoco, E., Cuccaro, P. and D'Onofrio, G. (2018). Simulation model for analysis and management of the no-show in outpatient clinic. Proceedings of the 10th International Conference on Computer Modeling and Simulation - ICCMS 2018. Sydney, Australia. 8-10 January 2018. ACM. pp. 236-241.

[13]. Alhamad, Z. (2013). 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. 2(2). pp.258.

[14]. Sheridan, R., Oliver, S., Hall, G., Allgar, V., Melling, P., Bolton, E., Atkin, K., Denton, D., Forbes, S., Green, T., Macleod, U. and Knapp, P. (2019). Patient non-attendance at urgent referral appointments for suspected cancer and its links to cancer diagnosis and one year mortality: A cohort study of patients referred on the two week wait pathway. Cancer Epidemiology, 63. p.101588.

[15]. Ding, X., Gellad, Z., Mather, C., Barth, P., Poon, E., Newman, M. and Goldstein, B. (2018). Designing risk prediction models for ambulatory no-shows across different specialties and clinics. Journal of the American Medical Informatics Association. 25(8). pp.924-930.

[16]. Devasahay, S., Karpagam, S. and Ma, N. (2017). Predicting appointment misses in hospitals using data analytics. mHealth. 3. pp.12-12.

[17]. Lee, G., Wang, S., Dipuro, F., Hou, J., Grover, P., Low, L., Liu, N. and Loke, C. (2017). Leveraging on predictive analytics to manage clinic no show and improve accessibility of care. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Tokyo, Japan,19-21 October 2018. IEEE. pp. 429-438.

[18]. Wolff, D., Waldorff, F., von Plessen, C., Mogensen, C., Sørensen, T., Houlind, K., Bogh, S. and Rubin, K. (2019). Rate and predictors for non-attendance of patients undergoing hospital outpatient treatment for chronic diseases: a register-based cohort study. BMC Health Services Research. 19(1), p.386.

[19]. Samorani, M. and LaGanga, L. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. European Journal of Operational Research. 240(1). pp.245-257.

[20]. George, A. and Rubin, G. (2003). Non-attendance in general practice: a systematic review and its implications for access to primary health care. Family Practice. 20(2). pp.178-184.
[21]. Douglas, E., Wardle, J., Massat, N. and Waller, J. (2015). Colposcopy attendance and deprivation: A retrospective analysis of 27 193 women in the NHS Cervical Screening Programme. British Journal of Cancer. 113(1). pp.119-122.
[22]. Gromisch, E., Turner, A., Leipertz, S., Beauvais, J. and Haselkorn, J. (2020). Who is not coming to clinic? A predictive model of excessive missed appointments in persons with multiple sclerosis. Multiple Sclerosis and Related Disorders. 38. p.101513.
[23]. Lenzi, H., Ben, Â.J. and Stein, A.T. (2019). Development and validation of a patient no-show predictive model at a primary care setting in Southern Brazil. PloS one. 14(4). p.e0214869.
[24]. Elvira, C., Ochoa, A., Gonzalvez, J. and Mochon, F. (2018). Machine-Learning-Based no show prediction in outpatient visits. International Journal of Interactive Multimedia and Artificial Intelligence. 4(7). pp.29-34.
[25]. Deceuninck, M., Fiems, D. and De Vuyst, S. (2018). Outpatient scheduling with unpunctual patients and no-shows. European Journal of Operational Research. 265(1). pp.195-207.
[26]. Alaeddini, A., Yang, K., Reeves, P. and Reddy, C. (2015). A hybrid prediction model for no-shows and cancellations of outpatient appointments. IIE Transactions on Healthcare Systems Engineering. 5(1). pp.14-32.
[27]. Topuz, K., Uner, H., Oztekin, A. and Yildirim, M. (2017). Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network. Annals of Operations Research. 263(1-2). pp.479-499.
[28]. Guy, R., Hocking, J., Wand, H., Stott, S., Ali, H. and Kaldor, J. (2011). How effective are short message service reminders at increasing clinic attendance? A meta-analysis and systematic review. Health Services Research. 47(2). pp.614-632.
[29]. Alyahya, M., Hijazi, H. and Nusairat, F. (2016). The effects of negative reinforcement on increasing patient adherence to appointments at King Abdullah University Hospital in Jordan. INQUIRY: The Journal of Health Care Organization, Provision, and Financing. 53. p.004695801666041.
[30]. Mohamed, K., Mustafa, A., Tahtamouni, S., Taha, E. and Hassan, R. (2016). A quality improvement project to reduce the 'no show' rate in a paediatric neurology clinic. BMJ Quality Improvement Reports. 5(1). pp.u209266.w3789.
[31]. Joni, H. (2017). Medical Appointment No Shows. [Online] Kaggle.com. Available from: <https://www.kaggle.com/joniarroba/noshowappointments?select=KaggleV2-May-2016.csv> [Accessed: 26 January 2021].
[32]. Witten, I., Frank, E., Hall, M. and Pal, C. (2017). Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques. 4th ed. Amsterdam: Morgan Kaufmann.
[33]. Zheng, A. and Casari, A. (2018). Feature Engineering for Machine Learning. 1st ed. Sebastopol: O'Reilly Media, Inc.
[34]. Sánchez-Hernández, F., Ballesteros-Herráez, J., S. Kraiem, M., Sánchez-Barba, M. and Moreno-García, M. (2019). Predictive modeling of ICU healthcare-associated infections from imbalanced data. Using ensembles and a clustering-based undersampling approach. Applied Sciences. 9(24). pp.5287.
[35]. Brownlee, J. (2020). Undersampling Algorithms for Imbalanced Classification. [Online] Machine Learning Mastery. Available from: <https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/> [Accessed: 5 February 2021].