No-Show Prediction for Medical Appointments

Rashmikant T Shukla  
*M.Sc. in Data Analytics*  
*National College of Ireland* Dublin, Ireland  
 x18181236@student.ncirl.ie

*Abstract*—This project design document presented application of predictive modelling in healthcare system. Health service provider all around the world bear reasonable loss due to no-show of patients. This document gives description of different applicable techniques and a strategy that will be employed to produce a predictive model in second phase of this project. It defines scope and goal based on the dataset also put the hypothesis and its probable evaluation methods.

Keywords—predictive model, evolution method, no-show.

# Background and Scope

Almost all appointment-based service providing industry faces problem of non-attendance which affect the whole system’s machinery. Similar problem is also faced by healthcare industry. Here patients’ non-shows are continuous worry of the healthcare providers. [1], defines non-shows as those cases where patients do not attend or cancel appointments in time so that it can be assigned to other patients. It results in insufficient use of healthcare personnel and other facilities this situation is very alarming in those regions where the demand of healthcare services is very high. There are many reports and articles from different countries reported a substantial economic loss in healthcare due to missed appointments.

In 2018 around 1300 people missed their outpatient appointments in Ireland due to which Health Service Executives struggles to manage the appointment waiting list [2]. In United Kingdom around 15 million general practice appointment are missed by patients every year[3]. According to NHS average cost of each appointment is around 30 pound that means around 216 million pounds are wasted each year. If one accounts the cost of inconvenience to other waiting patients and to medical staff, then it is quite alarming situation. [4], reports that cost of missing appointment in United State of America is 150 billion dollars each year the no shows rate is about 30% nationally. These few cases are of developed countries, when it comes to developing countries the numbers are more alarming and difficulties faced by patients are much more severe.

No-shows of patient to medical appointments varies across the globe states that Africa has around 43%, South America has around 27%, North America has around 23%, Europe has around 19%, Asia has around 25% of no-show rate [1]. So, coming up with generalize method to deal with this problem is difficult task but the specific region based, or locality-based solution are already in place which are working to improve the no-show rate of patients most of them revolve around scheduling of appointment.

## Scope

Appointment scheduling is in use from long time and there are various steps are included into it to improve the show rate of people. [5], discussed the different intervention like reminders, reduction in perceived barrier and increasing motivation, this paper also calculates effectiveness of these interventions. [6], Overbooking is also one of the used methods in appointment scheduling to tackle the no-shows. Nowadays, there are various applications are available to manage the appointment which already captures quite useful data which can be utilized in prediction. Use of predictive analytics will provide data driven intelligence to take more accurate intervention, overbooking. It will also provide data specific supports to these methods.

No-show of patient is a problem which is costing big monetary loss and non-efficient utilization resources. There are many technology-based solutions for scheduling of appointment combined with different notification techniques. Predictive analytics model will make appointment scheduling more intelligent. Overbooking decision taken by facility will be more accurate to mitigate the effect of no-shows.

## Dataset

[7], Dataset used for no-show prediction is taken from Kaggle data repository it is a publicly available dataset. It does not have the specific information regarding the method of data collection. This dataset is consisting 1,10,527 appointment records along with 14 variable which are maintained during appointments. Below are brief descriptions of each variables

* Patient ID – It is randomly generated number which is assigned to everyone to make unique identification.
* Appointment ID- structured sequence of number used while registering the appointment.
* Gender- Male (M) or Female (F)
* Scheduled Day- date on which appointment was created.
* Appointment Day- date of appointment
* Age- Age of the patient for whom appointment is created it ranges from 0 to 115
* Neighborhood- around 80 regions of brazil are present in the dataset
* Scholarship- have entries as 0 and 1 for not receiving and receiving scholarship respectively. Brazil government runs a scholarship scheme for economically backward people, in which they provide monetary aid to families, but the term is that if they have children, they should maintain pre-decided attendance in school.
* Then dataset have few common medical records viz. Hypertension, Diabetes, Alcoholism, Handicap these all are 0 and 1 values.
* SMS Received- if the patient received appointment information SMS or not.
* No-Show- it is our target variable which is in form of 0 or 1 for attended the appointment and not attended the appointment respectively.

# Goals of the Project

The main goal of this project is to predict no-show of patients based on their previous appointment data using appropriate machine learning technique. This will help to mitigate no-show problem and will enable healthcare service providers to use resources more efficiently and reduce the monetary losses. Which will enable them to provide services to needful.

Along with this, presented paper also tries to find different pattern and trends like

* Proportion of male and female missing the appointment?
* Is there a trend between different age group and no-show or Combined effect of gender and age on no-shows?
* Is there any day on which missed appointment are comparatively very high?
* Relationship between no-shows and other captured data like diabetes, scholarship, handicap, hypertension, alcoholism?
* What is the relationship between different regions of brazil and show and no-shows?
* What is the prominent factor affecting the no-shows of patients?

# Ethical Concerns

When dealing with the medical data there are many ethical and operational concern which one must address before using the data. Usually predictive analytics modelers are recommended to use data which is already been collected without any explicit consent from the patients, but the model developer should follow federal law regarding privacy of patients [8]. In other words, model developer should not have access to any data from which they can trace back the identity or any critical information about the patient. But in few cases like in study of genome data it is difficult to ensure above mentioned requirements.

[9], talks about a model which can be used by data collectors in health care. It should be decided before hand which organization is going to use this data, is there any extra burden or risk on participant. Participants should be aware of the benefit of it to themselves and as well as society.

Data used in this paper is available publicly it was originally generated by the research authority in Brazil for the study purposes. It does not have any information with whom individual identity can be discovered. So, it does not raise any ethical concerns. This dataset only represents a sample of patients who missed their appointments in Brazil and few of their health indicators like diabetes, hypertension, alcoholism.

The proposed method does not try to find out any such thing which can be potentially breach the privacy of any individual. In all situation it follows all the GDPR instructions. Dataset used, has customer id column, which is randomly generated numbers, to ensure the privacy of everyone.

# Strategy of Analysis of dataset

Strategy for this project started with the selection of research topic and appropriate dataset of selected topic. Before finalizing the research topic, it was discussed with the professor and after getting nod from professor we moved to next phase.

## Dataset Selection

dataset is chosen from a publicly accessible data repository.

## Recognition of dependent and independent variable

Dataset has 14 variables out of which No-show column is the Dependent variable and other 13 variables are independent one. Selected dependent variable is categorical with two levels No and Yes.

## Hypothesis

* H0 = Proportion of No-show of patient to their medical appointments predicted by the selected independent variables <= 50%
* H1 = Proportion of No-show of patient to their medical appointment predicted by selected independent variables > 50%

Confidence level of this analysis is 95% or the selected p-value is 5%.

## Methodology

Analysis in this paper will follow Knowledge Discovery in Databases (KDD) technique. KDD is one of preferred methodology used by researcher in machine learning and predictive analytics. It leverages the researcher from including domain expert in the selected field. The question this paper addressing is very much a business case study that mean here one can employ Cross Industry Standard Process for Data Mining (CRISP-DM) Since this paper does not studies the deployment and we do not have industry expertise of healthcare service industry. That’s why preferred methodology is KDD.

## Variable which are out of scope in Presented Analysis

This research focuses on predicting the no-show by limited set of explanatory variables, there are other factors which can be used improve the final model. Few of them are discussed below

* Income Group is one of the factors which is not included in this study as there is lack of public income data for the selected group of samples. Individual or family Income may have effect on attendance to medical appointments
* Travelling Options also one of the variables can be used in this type of problems but this study does not cover it.
* Insurance Status is also one of the factors which affect medical services which is not covered by presented study.
* Education level might also have effect on the no-show prediction as education is resemblance of awareness regarding medical facilities.
* Distribution of medical facility is also one of the factors which is not considered into this study.

This paper also provides a section on applicable predictive analytics techniques based on available literature. In next phase, one of the suitable techniques will be chosen and result will be presented.

## Defining the Classification Matrix

As the target variable is dichotomous and there is imbalance in target variable. No-show cases is comparatively lower than the non-missed appointment. So, there is always chances of model being biased towards the non-missed cases, but the focus is on reducing no show cases.

* True Positive Count- are total number of people which attended the appointment predicted correctly by proposed model.
* False Positive Count – are those which attended the appointment, but model predicted them as no-show.
* False Negative count- are those which are no-show cases, but model predicted them as people who attended the appointment.
* True Negative Count- are the count of correctly predicted no-show by model.

There is always trade-off between True positive rate/sensitivity and True negative rate/specificity. In this problem focus is on model which is more robust to giving high Specificity.

## Applicable Evaluation Methods

As target variable is categorical so we end up with classifying problem so for evaluation classification table can be used. From this different metric like accuracy sensitivity and specificity can be used. AUC and ROC curve can be used to evaluate the result.

# Preliminary Visualizations

Visualization are done using excel and python. In python there are various libraries are available in this matplotlib seaborn libraries are used.

### Target variable distribution

Fig:1

Fig 1 shows that target variable is imbalanced it have 20% cases of missed appointment and remaining 80% are showed up to their appointments.

### Gender based no-show distribution

Fig:2

Form Fig 2 shows the gender-based distribution of target variable. Overall representation of female is more than male. In both show and no-show cases females are almost double to that of male. We future investigate whether there is any trend based on different age groups in male and female both.

### Age and Gender-wise Appointment Trend

From trend in Fig 3 one can conclude that till age of 15 male and female both visit doctors almost in similar pattern but after that these changes drastically between age of 15 to 60. Females visit doctors more often in their fifties. After 60 trends steeply declined for male and female both.

Fig:3

### Trend of waiting time

Fig:4

Above trends suggest that in dataset as the waiting period is increased the chances of not attending the appointment is decreases drastically.

More people showed up for their appointments is the wait period was below 25 days even in that rate of “show” are very high for 10 days waiting period.

### Day based no-show

Fig:5

There is not much trend based on different days of week. From this trend seems like people prefer Monday and Tuesday for the appointments.

### Notification-based Trends

Fig:6

From this trend it is quite interesting to see that even if notification is send via SMS about appointment there is no considerable change in “no-show” numbers.

### Different condition-based Trends

Fig:7

Above figure tells that diabetes patients are quite frequent with their appointments and their show count is also very high.

Fig:8

Figure 8 compares count of no-show in different conditions which are present in dataset.

No-show count is high in people with hypertension.

After receiving SMS there is no positive effect in people and their no-show count is high.

Fig:9

If we plot smoker’s data against the appointment and age, then trend shows that population in Brazil smokes more in their 40’s. Most of them start in their late 20’s which in line with normal known trend that during college time most people start smoking.

As they grow and come to retirement age number of people smoking decreases.

In dataset there is a peak at 100 year which might have occurred due to error or it can be one of the outliers.

### Month-wise Trend

Fig:10

If we plot appointment based against month then may month have received highest appointment in the dataset and no-show count is also high, but this can also be due to large count of appointments were done in May. So, no-show data is also high.

April and June months have approximately similar trends.

# Applicable Techniques

Based on literature available on techniques used in this field one can say the focused is on model which can be explained well, less use of complex model is prevalent, but in application-based systems even complex model are employed.

* Different Combination of Logistic Regression Models

[6], uses stepwise naïve and mixed effect logistic regression model. This uses Akaike Information Criteria (AIC) to select best model. Selected best model was evaluated using Area under the curve (AUC). In this study same day appointment and the previous record of attendance are comes out to be most important predictor. [10], also uses logistic regression in three different clinic and compare the result of these. In all three around 62% to 65% were predicted correctly.

[11], also uses multivariable logistic regression but here chi square and t-test are used prior to applying the model. From these tests, variables whose p-value is more than 0.25 are selected for the final model.

* Different Tree based Models

[10], uses decision tree model along with cross validation to predict the “Show” and “No-Show”, for the evaluation it uses ROC curve. There are other Tree methods like Random Forest or Tree method with bagging boosting techniques can be employed as our dataset have imbalance of class in target variable. It is a good choice as in our case expansibility of model is important but decision tree are very sensitive to data on which they are applied.

* Optimization Approach

[12], address the problem of no-show prediction using Association rule mining and optimization technique. This paper derived three rules which are combination of around 3 or 4 explanatory variable and in the end by simulation prove that applying these rules they received more accuracy in predicting no-show of patients.

* Probability based Models

[13], uses Bayesian analysis, in which Markov Chain Monte Carlo is used to find posterior probability. Then this result is fed into final predictive model.

[14], take one clinic and built a hybrid model on their existing appointment scheduling system. In this Markov Decision Process is applied in collaboration with overbooking. This study recognizes one threshold for clinic which can be used in taking overbooking decisions.

In present a probabilistic Naïve model where for each person’s Likelihood Ratio (LR) calculated for “show” and “No-show” [15]. It also uses ROC for evaluation and cross validation to select final model.

* Neural Networks

There are other papers of appointment no-show in different field like in air ticket booking system. [16], discussed the implementation of neural network in the no- show prediction.

As it is black box kind of model its interpretation is very difficult. So, there is limitation regarding business explanation.

##### References

[1] L. F. Dantas, J. L. Fleck, F. L. Cyrino Oliveira, and S. Hamacher, “No-shows in appointment scheduling – a systematic literature review,” *Health Policy*. 2018.

[2] E. Ring, “479,000 outpatient no-shows last year,” *Irish Examiner Reporter*. [Online]. Available: https://www.irishexaminer.com/ireland/479000-outpatient-no-shows-last-year-466843.html. [Accessed: 08-Mar-2020].

[3] NHS, “NHS England » Missed GP appointments costing NHS millions,” *NHS*. [Online]. Available: https://www.england.nhs.uk/2019/01/missed-gp-appointments-costing-nhs-millions/. [Accessed: 08-Mar-2020].

[4] J. Gier, “Assure smart, complete scheduling to maximize capacity and utilization.”

[5] A. George and G. Rubin, “Non-attendance in general practice: A systematic review and its implications for access to primary health care,” *Fam. Pract.*, 2003.

[6] H. Lenzi, Â. J. Ben, and A. T. Stein, “Development and validation of a patient no-show predictive model at a primary care setting in Southern Brazil,” *PLoS One*, 2019.

[7] Kaggle, “Medical Appointment No Shows | Kaggle,” *Kaggle*. [Online]. Available: https://www.kaggle.com/joniarroba/noshowappointments#KaggleV2-May-2016.csv. [Accessed: 08-Mar-2020].

[8] I. G. Cohen, R. Amarasingham, A. Shah, B. Xie, and B. Lo, “The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care,” *Health Aff.*, vol. 33, no. 7, pp. 1139–1147, Jul. 2014.

[9] D. Wade, “Ethics of collecting and using healthcare data,” *British Medical Journal*. 2007.

[10] N. L. Ma, S. Khataniar, D. Wu, and S. S. Y. Ng, “Predictive analytics for outpatient appointments,” in *ICISA 2014 - 2014 5th International Conference on Information Science and Applications*, 2014.

[11] J. Daggy *et al.*, “Using no-show modeling to improve clinic performance,” *Health Informatics J.*, 2010.

[12] K. J. Glowacka, R. M. Henry, and J. H. May, “A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling,” *J. Oper. Res. Soc.*, vol. 60, no. 8, pp. 1056–1068, Aug. 2009.

[13] S. M. O’Brien and D. B. Dunson, “Bayesian Multivariate Logistic Regression,” *Biometrics*, vol. 60, no. 3, pp. 739–746, Sep. 2004.

[14] P. Cao and J. Tang, “Hybrid appointment for a single-physician clinic with no-shows and overbooking,” in *26th Chinese Control and Decision Conference, CCDC 2014*, 2014.

[15] V. Levy, “A predictive tool for nonattendance at a specialty clinic: An application of multivariate probabilistic big data analytics,” in *2013 10th International Conference and Expo on Emerging Technologies for a Smarter World (CEWIT)*, 2013, pp. 1–4.

[16] R. Z. Cao, W. Ding, X. Y. He, and H. Zhang, “Data mining techniques to improve no-show forecasting,” in *Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics*, 2010, pp. 40–45.