No-Show Prediction for Medical Appointments

Rashmikant T Shukla  
*M.Sc. in Data Analytics*  
*National College of Ireland* Dublin, Ireland  
 [x18181236@student.ncirl.ie](mailto:x18181236@student.ncirl.ie)

*Abstract*— No-show of patients to their medical appointments causes heavy monetary loss to health care systems all around the world. Other than economic losses it disrupts the whole system. Traditionally to deal with this, methods like overbooking and giving reminder are used but to make this more efficient there have been lots of research to develop machine learning prediction algorithm for no-show predictions. Although these models differ for different medical setting, but prediction makes appointment system more intelligent. This paper presents the application of tree based Random Forest technique for prediction of no-show of patients to their medical appointment. The performance of this model is compared with the decision tree model from the literature. Comparison is done on basis of accuracy and recall. Presented analysis is performed on the dataset presenting appointment details of Brazil’s healthcare system.

Keywords—Random Forest, Recall, Decision Tree, No-Show.

# Introduction

This document is for the second phase of this project. The phase one document had introduced the topic under consideration for this predictive analytics project and discussed how it is relevant problem and how predictive analytics can come to recue here. Then in next section it stated the goal of the project in detail followed by section on ethical concerns, as the selected dataset is publicly available there is no issues regarding this. The next section of the document provided detailed discussion on strategy by hypothesising the problem and by providing brief accounts on methodology, out of scope variables, classification matrix, evaluation methods. The next section of the document presented initial visualisation on selected dataset. After that last section of document gave a brief literature review by keeping focus on different kind of machine learning techniques used for similar kind of problem.

No-show of patients on their medical appointment is very concerning for healthcare system and its effective management become more important in countries who already suffers from lack of effective system and lack of medical staff. This no-show of patients put extra burden on system and there is always chances of non-efficient use of medical practitioner along with considerable monetary loses. Medical system all over the world faces 19% to around 43% of no-show of patients [1]. Europe region being the lowest and Africa region being the highest in no-show rates.

Different methods are used by medical authorities to tackle the no-show problem like reminders, motivating the patient, overbooking to account for the no-shows but count for overbooking needs to be precise if it is less then doctors may have free time and if it is more then it will result in more waiting time, causing dissatisfaction in patients. So, machine learning prediction is effective mean to arrive at the perfect count of overbooking or in many cases result can be very accurate for certain patients’ probability of attending the appointment. So, this research will help authorities to tackle the no-show more effectively.

## Hypothesis

H0 = Selected set of independent variables can predict no-show proportion in medical appointment by less than 50% accuracy.

H1= Selected set of independent variables can predict no-show proportion in medical appointments by more than 50% accuracy.

Confidence level of the analysis is 95%.

In second phase we moved to implantation part of the selected technique for the prediction of no-show of people to their medical appointments using machine leaning techniques. In next section this paper gives a brief account on the Related Work in this field followed by sections on Methodology and Implementation in which under various subsections Target dataset, Pre-processing, Feature Engineering, correlation matrix and Algorithm Selection & model creation are discussed. In next section result and conclusion is presented.

# Related Work

In literature many machine learning techniques are used to predict the no-show of patients. In phase one also we have discussed different machine learning techniques used in similar problems. More focus being on the techniques with more interpretability like logistic regression, different decision-based techniques.

[2] worked on a Brazilian healthcare system’s appointment data, they presented stepwise naïve and mixed effect logistic regression. This dataset is similar to, the dataset used for this prediction project. Here the focus is on interpretation so in order to reduce the complexity logistic regression is perfect choice it is very simple and effective technique where one can see the effect of each variable on the target variable. It establishes the logit linear kind of relation between dependent and independent variables. They use 50% of dataset for training and 50% dataset for testing which is very unlikely as in machine learning project usually data division is around 70-30% or more. This may give less data for the model to train. Akaike Inflection Criteria (AIC) is used to select best performing model in naïve logistic regression approach. They have used mixed effect logistic regression since the data have many categorical variables and all this variable do not contain all possible outcome that is, they are not complete representation of population so mixed effect logistic regression technique is good selection. Mixed effect logistic regression model had the low AIC value, so it is selected as final model and father evolution is done using Area Under the Curve (AUC) and sensitivity of the model. The model has achieved 81% accuracy. From this model it is found that previous appointment and same day appointment are best predictor for this problem. Mixed effect logistic regression takes less account of baseline predictor which is a limitation in some cases. As it is binary classification type of problem even tree-based technique can give good result and high level of interpretation.[3] also used Logistic Regression for no-show prediction but it presented two other parts; first finding the individual probability of cancellation or no-show based on individual history for this Bayesian techniques was used, second assigning weights to different factors. In the end, result from these three approaches was integrated to give final model. Final model’s Mean Square Error (MSE) was compared with other method and it performed better than others. Receiver Operating Curve (ROC) is also used for the evolution of the model. This also shows the focus is on interpretability of model.

[4] presented a study on three model logistic regression, Naïve Bayes and multilayer perceptron. Used data was from Indianapolis, logistic regression model used stepwise selection method. Naïve Bayes and Logistic Regression performed well and visit type, age, insurance was the important variable. When compared the AUC of different model Naïve Bayes classifier had 0.86 which is highest in comparison to 0.81 and 0.66 in logistic regression and perceptron model respectively. This study doesn’t comment on sensitivity which is one of the major factors to evaluate these kinds of problems. Although performance of naïve Bayes is better, but it does not take account the interaction effect of variables.

There are various probabilistic models which are used by researchers to address this problem. [5] [6]uses Markov chain to find the probability of no-show and this result along with overbooking gives good results. In other instant posterior probability result used as input into final algorithm to make more accurate predictions.

[7] have worked on 10 years of data of one clinic and used logistic regression for this, the goal of the analysis was to predict no-show of patients and based on that supplement the existing overbooking technique. Overbooking [8] is one of the techniques which is commonly used in appointment-based setting to tackle the problem of no-show. In this usually extra number of appointments are done by hoping that some of them will not turn up for the appointment. Here if the extra appointment done are not accurate then it may cause disruption in services. This paper compares overbooking with prediction of no-show, random overbooking and evenly distributed overbooking and assessment is done on factors like patient wait time, medical practitioner average idle time, overtime in a day and performance of prediction-based overbooking is better. This shows the use of prediction in this kind of problem which will make existing system more intelligent.

As discussed, average idle time, overtime and wait time can be used to calculate expected cost so this can contribute in optimization. So, problem no-show of patients can also be solved by using optimization techniques.[9] uses predictive analytics along with optimization techniques to optimize the overbooking. In this prediction is done for individuals and then this result is used in optimization. This is an effective technique, but it is very specific to clinic setting there are very few factors which can be used in general no-show predictions.

[10] used a different approach to tackle the problem of no-show. They applied Association Rule Mining technique to derive rules to overcome the problem of no-show. Although rule-based method is easy to explain but usually very large number of rules are derived and they become quite confusing to distinguish and they are very case specific, generalization is difficult for continuous independent variables generalized rule induction routine are followed. Then these rules are used in simulation model to make scheduling systems more efficient.

Other than this there are papers which uses support vector machine, neural networks for no-show prediction[4].[11] also uses deep learning method for no-show prediction. Although the results are very competitive, but the lack of explanation is big drawback. These models don’t give relation between different dependent and independent variable or combined effect of different independent variable. Understanding these relationships are important to make decision regarding reduction of no-show of patients.

As priority is interpretation of models for this tree-based machine learning methods are quite popular.[12] uses decision tree model with cross validation for no-show prediction in outpatients. Then ROC and lift curve was used to see quality of this model further misclassification rate is calculated to assess the performance of the model.

Similarly [13] also uses Decision Tree technique for the prediction. For this used dataset is same as of this research that is around 100k appointments record of multiple regions in Brazil. They have provided a detailed univariate bivariate and multivariate analysis of the independent variables. They also provide brief accounts on recurring patients and new patients how it affected the no-show of the patients. For the analysis they have used logistic regression and decision tree algorithms. Logistic regression model achieved accuracy of 86 % and precision 72% and recall of 50%. On the same data Decision Tree achieved 89% accuracy, 74% precision and 73% recall. So, decision tree has comparatively high accuracy and shows significant improvement in precision and recall. This paper also uses same dataset and these results shows that tree-based technique is performing better, based on this a random forest algorithm is used for this analysis. Usually Random Forest algorithm are better in generalizing than decision tree and even they are more efficient with unseen data, while decision trees are quite sensitive to change in dataset.

Overall literatures show there are mixed of simple and complex techniques are used for this problem. As focus of this paper is on interpretation complex methods like perceptron model and deep learning will not be feasible. When it comes to simple algorithm with high interpretability, literature shows application of different form of logistic regression models. It is also evident since the problem in hand is binary classification many have gone with logistic regression. Then there are papers who have used decision tree this method has improved the performance of models. It also supports the intent of using random forest technique for no-show prediction of medical appointments.

# Methodology and Implementation

As mentioned in first phase of the project analysis will be done according to Knowledge Discovery in Dataset (KDD) methodology. This methodology is popularly used for data mining or data science projects. As in earlier sections the domain of the research and its application are discussed in detail, we moved to the data selection part. All the analysis in coming sections are done in excel and Jupyter notebook is used.

## Target Dataset

For this project dataset is taken from online data repository Kaggle. This is an open dataset with 1,10,527 records, it has 14 variables. These variables are categorical and continuous in nature. This dataset originally came from Brazil’s healthcare system. Data source does not have any account on data collection process. Out of 14 variables no-show column is dependent variable it is binary in nature with entry as YES and NO. below table give the data description

TABLE I: Data Description

|  |  |  |
| --- | --- | --- |
| ***Variable Name*** | ***Descriptions*** | ***Type*** |
| Patient ID | Randomly assigned number by system | Numerical |
| Appointment ID | Structured sequential number assigned to each patient | Numerical |
| Gender | Male or Female | Categorical |
| Scheduled Date | Day on which appointment was registered | Date |
| Appointment Date | Day of appointment | Date |
| Age | Age of patients | Numerical |
| Neighbourhood | Different regions | Categorical |
| Scholarship | Brazil government scheme | Categorical |
| Hypertension | Medical Condition | Categorical |
| Diabetes | Medical Condition | Categorical |
| Alcoholism | Medical Condition | Categorical |
| Handicap | Physical condition | Categorical |
| SMS\_received | notification received | Categorical |
| No-show | Person attended the appointment or not | Categorical |

## Pre-processing

At this stage data was uploaded into python and preprocessing and initial work on data started.

### Checks for Missing value

The selected dataset does not have any missing value. In next section we will check each variable if any irregularities are seen and if there is any then it will be updated accordingly.

### Univariate Analysis

In univariate analysis each variable is analyzed individually. If there is some inconsistency, then it will be updated. Few variables are discussed below on which major changes are done.

#### PatientID

This variable was having float datatype, so it is converted into integer and this variable is used to find out distinct patients. Data set have 110,527 appointment out of which 62,299 are from distinct patient. That means around 40% patients have more than one appointment. This information can be used to create previous appointment records. This type of variable was used in [2].

#### Scheduled date and Appointment date

These two variables are converted into standard date format individually they do not have much use in model building but using these two new variable can be created like days between the schedule date and appointment day as few of researcher have pointed out that if gap is more, then chances of not coming for the appointment are more. So, this will be checked in Feature engineering phase.

#### Age

In literature age was one of the important variables in selected dataset there was one negative entry which is probably an error during data collection. So, it was removed. One interesting thing this variable had around 3500 entry as zero. Initial thought was that it could be mistake or babies below 1-year age, to confirm this corresponding column of hypertension and alcoholism was checked all the corresponding entry were 0. Since one knows these conditions are not seen in babies. So, conclusion was that dataset have entries of babies below 1-year. This column has some outliers also like few entries have 115 value, but these are kept as it is.

#### Handicap

This variable has 5 different entries so to make it more sensible entries were converted into 0 and 1.

1. Person without handicap status.
2. Person with handicap condition.

#### No-show

This is our target variable it has entries as YES and No. it is converted into binary form

1. Person who came to their appointment
2. Person who missed the appointment

Other than this there was no major issues with variable like Scholarship, Hypertension, Alcoholism, Diabetes, SMS\_received etc.

## Feature Engineering

In previous stage we have processed the data, based on that patient ID and Appointment ID are dropped as they do not have any useful info which can contribute to the final model. As per literature weekdays, weekend and time between scheduling day and appointment day are important variable. So, we will calculate these variables from existing data. Below are the new variables generated from existing data.

### DaysBeforeApp

By using Scheduled date and appointment date new variable ‘DaysBeforeApp’ is calculate it shows that how many days in advance appointment was created. Also, literature shows longer the duration between schedule date and appointment more is the chance of missing the appointment.

* While calculating days before appointment variable there were 6 negative value. In real it is not possible probably this is a mistake. So, these observations are deleted as the number are small it would not have any effect. If large observation were negative, then may be replacement option was chosen.
* Around 38000 people have same scheduled date and appointment date.
* To balance this variable, it further converted into different bins.
* 0-days, 1-2 days, 3-7 days, 8-31 days, more than 31 days. Below figure gives the count of each bins.

Fig 1. Distribution of Bins

### Weekday Appointment

Based on appointment date variable all appointments are segregated into different days of the week.

Fig 2. Appointment by Weekdays

### PreviousApp

This variable is created based on literature which shows that person who have booked appointment previously have less chance of missing the appointment.

## Correlation Matrix

In previous section three variable are generated which will be used in final model creation before model creation correlation is checked to make sure there is no multicollinearity it will make sure there is no redundant variable in model building process. Below is the heatmap of correlation matrix.

Fig 3. Correlation Heatmap

From figure 2, it’s clear that there is no significant correlation between the independent variable. So, we are good to go for next stage of the analysis.

## Algorithm Selection and Model Creation

[13] has worked on same dataset as of this paper it shows that decision tree method performed better than logistic regression which is one of the most used technique for this problem.

* For classification type of problems Decision tree and random forest are used but random forest outperforms decision tree on many occasions. Random forest has multiple trees and each split is performed randomly which reduce the overfitting [14].
* In decision tree mostly trees are pruned but in Random forest tree are allowed to grow fully so it gives better performance [15]
* As in random forest random sample is used so tree is diverse in nature.
* Although random forest losses the interpretability but feature importance method can be used to find the important feature for the model.

Due to all these benefit over decision tree and logistic regression Random forest algorithm is used for the final model creation. Now that algorithm is selected, we moved to next stage of the project.

### Data Partition

Before model creation total dataset is divided into 2 part one having 70% of data and other having 30% of data. These two set are used for training and testing of the model.

### Model Creation

Below image shows the classifier used in the model creation.

Fig 4. Random Forest Classifier

# Result and Conclusion

## Model Evaluation

In this section evaluation of implemented Random forest is presented. For evolution of model confusion matrix, ROC curve and Recall and precision is used.

* Confusion matrix is presented below

Fig 5. Confusion Matrix

From confusion matrix it is clear that random forest model performed quite well and predicted majority of show and no-show correctly. 93% are classified correctly.

* ROC curve

ROC curve shows that model have achieved very good precision and recall. Recall on testing data is 82.6%.

Fig 6. ROC Curve

If the performance of this model is compared with the model presented by [13] then random forest model achieved more accuracy, recall than Decision Tree model.

## Feature Imporatnce

Feature importance is performed for the random forest classifier. Previous appointment, Same day scheduled, and appointment day, Age and notification are important variables for the prediction of no-show.

## Interpretation of Results

* So proposed model can predict 93% of cases correctly that mean we reject H0.
* Proposed model achieved recall off 82.6% that means 82% of relevant cases are correctly classified.
* Model committed very few Type-1 error.
* For this dataset Random forest can successfully classify most of the cases.
* Analysis shows that person who have taken the appointment on same day have less chance of missing the appointments.
* In model prediction previous appointment variable is very important which proves that decision of feature engineering was accurate and resulted into positive result.
* These result shows that this classifier can be used to make improvements on the existing appointment creation systems.
* Also, the decision on overbooking to accommodate the no-show of patients can be more precise.

# 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*, vol. 122, no. 4, pp. 412–421, Apr. 2018, doi: 10.1016/j.healthpol.2018.02.002.

[2] 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*, vol. 14, no. 4, p. e0214869, Apr. 2019, doi: 10.1371/journal.pone.0214869.

[3] A. Alaeddini, K. Yang, P. Reeves, and C. K. Reddy, “A hybrid prediction model for no-shows and cancellations of outpatient appointments,” *IIE Trans. Healthc. Syst. Eng.*, vol. 5, no. 1, pp. 14–32, Jan. 2015, doi: 10.1080/19488300.2014.993006.

[4] I. Mohammadi, H. Wu, A. Turkcan, T. Toscos, and B. N. Doebbeling, “Data Analytics and Modeling for Appointment No-show in Community Health Centers,” *J. Prim. Care Community Health*, vol. 9, p. 215013271881169, Jan. 2018, doi: 10.1177/2150132718811692.

[5] S. M. O’Brien and D. B. Dunson, “Bayesian Multivariate Logistic Regression,” *Biometrics*, vol. 60, no. 3, pp. 739–746, Sep. 2004, doi: 10.1111/j.0006-341X.2004.00224.x.

[6] P. Cao and J. Tang, “Hybrid appointment for a single-physician clinic with no-shows and overbooking,” in *The 26th Chinese Control and Decision Conference (2014 CCDC)*, Changsha, China, May 2014, pp. 2680–2685, doi: 10.1109/CCDC.2014.6852626.

[7] Y. Huang and D. A. Hanauer, “Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach,” *Appl. Clin. Inform.*, vol. 05, no. 03, pp. 836–860, 2014, doi: 10.4338/ACI-2014-04-RA-0026.

[8] L. R. LaGanga and S. R. Lawrence, “Clinic Overbooking to Improve Patient Access and Increase Provider Productivity,” *Decis. Sci.*, vol. 38, no. 2, pp. 251–276, May 2007, doi: 10.1111/j.1540-5915.2007.00158.x.

[9] M. Samorani and L. R. LaGanga, “Outpatient appointment scheduling given individual day-dependent no-show predictions,” *Eur. J. Oper. Res.*, vol. 240, no. 1, pp. 245–257, Jan. 2015, doi: 10.1016/j.ejor.2014.06.034.

[10] 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, doi: 10.1057/jors.2008.177.

[11] S. Alshaya, A. McCarren, and A. Al-Rasheed, “Predicting No-show Medical Appointments Using Machine Learning,” in *Advances in Data Science, Cyber Security and IT Applications*, vol. 1097, A. Alfaries, H. Mengash, A. Yasar, and E. Shakshuki, Eds. Cham: Springer International Publishing, 2019, pp. 211–223.

[12] N. L. Ma, S. Khataniar, D. Wu, and S. S. Y. Ng, “Predictive Analytics for Outpatient Appointments,” in *2014 International Conference on Information Science & Applications (ICISA)*, Seoul, South Korea, May 2014, pp. 1–4, doi: 10.1109/ICISA.2014.6847449.

[13] M. D. A. Praveena, J. S. Krupa, and S. SaiPreethi, “Statistical Analysis Of Medical Appointments Using Decision Tree,” in *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, Mar. 2019, pp. 59–64, doi: 10.1109/ICONSTEM.2019.8918766.

[14] A. S. More and D. P. Rana, “Review of random forest classification techniques to resolve data imbalance,” in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Aurangabad, Oct. 2017, pp. 72–78, doi: 10.1109/ICISIM.2017.8122151.

[15] P. H. Swain and H. Hauska, “The decision tree classifier: Design and potential,” *IEEE Trans. Geosci. Electron.*, vol. 15, no. 3, pp. 142–147, Jul. 1977, doi: 10.1109/TGE.1977.6498972.