Data Analytics and comparison of various models for predicting appointment No-Show.

Health Care System- BRAZIL

Bala Tripura Sundari Padavala
-*Huether school of Business: The College of Saint Rose, Albany, New York*
Nichita Jamwale
-*Huether school of Business: The College of Saint Rose, Albany, New York*

ARTICLE INFO

Submission date: April 7 2020

Keywords:
No-Show
Random under Sampling
Cross-validation
Brazilian Health system
Retrospective study
Bayesian Belief
Artificial Neural Network

ABSTRACT

Patient No-show is a prevailing issue in the healthcare service units guiding to allocate the inefficient resources and limited access to healthcare providers. No-show is defined as a scenario where a patient has an appointment and voluntarily/involuntarily does not show up for the appointment. These are problematic to healthcare systems at all levels. This study aims to compare various models in predicting whether a patient will show up/doesn't show up on his last appointment by training the data containing the information about his prior visits. It also aims to reveal various factors that would impact the no-show rate. A retrospective study was accomplished using appointments from the Brazilian Public Health System (2016). All the models were refined with 10-fold cross-validation and were under sampled using the Random-Under Sampling method. The AUC (ROC curve) was utilized to evaluate the best performance of the models. Of the 62,299 unique patients, 38,999 patients were correctly classified by Bayesian Belief and 37,651 patients were correctly classified by Artificial Neural Network with AUC of 74.2% each. The most significant factors impacting the no-show were wait-time, prior-appointments, pa_pn (Interactive variable between prior no-shows and prior appointments). It is believed that this study would be a useful reference for decision making in the scheduling system in any appointment-based units.

Contents:

MOTIVATION

Healthcare Providers have few compliant challenges to face in the real world that back them off from providing services based on value and revenue. No-show of patients being one of the scenarios that hinders them from treating the patients with quality. They create a great confusion for the health care systems and create trouble in maintaining better productivity levels. This paper segmented with usage of various data mining techniques as well to evaluate various models with the significant factors affecting the No-show of Patients.

RELATED RESEARCH

In Europe, the waiting time for the patients to meet the healthcare providers has been an issue from the perspective of community/society (Worthington 1987). The researchers in the United states have started to focus on waiting time of the patients as it started impacting the profits of the health care centres. Past researches have been mostly about estimating or finding the amount of no-show in the healthcare units. Lately another element has been added to it i,e., estimating not just estimating the amount of no-show but also the ways to lower its occurrence's. The literature has also shown that some other factors impacting the no-show would also be socio-demographic factors, income-levels of patients, educational background of the patients (Murdock Rodgers 2003). Overbooking was another solution suggested by few studies where a stochastic model was considered, and walk-in rates were applied. On the contrary this reduced the quality services provided by the clinics as the wait-time gradually increased due to higher number of appointments per day. Patients thought of avoiding the appointment to get rid of higher wait-times. Less data is one of the significant reasons for no-show in the private sector. Patient's lack of understanding about the health-care system could also lead them to not turning up for their appointment. It might also impact on the type of insurance the patients obtain.

Missed appointments mainly wastes patient's as well as patients time and it would hinder the sick people from taking the opportunity to get the services. The evolution of the fee system in the United States has really helped in improving operational efficiency to compensate the rising prices of the healthcare centres (Feldstein,1996). Also, Patient's experience starts when they could schedule an appointment when they need (time factor), their ability to get one, this would continue till they complete an appointment and finally lasts till completion of their plan. A study in the United Kingdom estimated the annual financial cost to be 790million pounds i.e.,(963.6 million dollars).Other factors like logistics, parking could also have a huge impact in making up for an appointment.

A study showed a positive correlation between a doctor's empathy and attentiveness, also patient's capacity to handle their situation (Zachariae al,2003). The older age factor was found in one of the previous researches to be a significant factor associated in missing an appointment. Seasons/Weather, Language Proficiency was also linked to the studies representing the factors for this study

INTRODUCTION

Incomplete appointments, or patient no-shows, are scheduled appointments that patients either do not keep or do not cancel in time for another patient to be scheduled as a replacement. The patient's non-attendance could compromise the core principles of primary care: The accessibility and the continuity of care [GeorgeA,Rubin (2003)]. Missed appointments reduce the efficiency of the health care system and negatively impact access to care for all patients. Identifying patients at risk for missing an appointment could help health care systems and providers better target interventions to reduce patient no-shows. Delay in the tests conducted in the Health care centres could potentially put patients in danger, when a patient misses a screening or doesn't show for the appointment, he/she mad delay their treatment or in detecting a disease. thus, reduce in the No-Show rate not only diminishes the cost but also improves the quality of delivering services to the patients. This term was familiarized from the airline companies and later was used by hotels and healthcare sector. Seven features in the health care sector determine its quality [Alessandra Trindade Machado]:

1. Efficacy- Refers to one's belief in providing care and improving patient's health.
2. Efficiency- Refers to avoiding wastage of resources and efforts of the healthcare providers.

3. Effectiveness- Refers to practical impact of efficacy and how well the clinical trials would perform.
4.Acceptability-Refers to adjustments made for necessities & preferences of patients with care.
5.Optimality- Refers to analysing best way of cost-reduction and well-being of Patients.
6. Legitimacy- Refers to better cooperation and Community relations in terms of health services offered.
7.Equity- Refers to the belief of working together as one and all and all for one.

Efficiency is also stated to be a legal norm in the Constitution from the Brazilian Constitutional Amendment (EC) n.19/98 with the call Administrative Reform, Reform of Public Management and Managerial State Reform [Alessandra Trindade Machado]. The frequency of no-show is spread all over the world, but the majority is observed in the developing countries as well as in low income countries. There has been very little data on this issue from Brazil, still it has been reported that there has been about 48.9% of no shows at the level of Primary care [Henry Lenzi,Angela Jornada Ben]. There have been several studies suggesting to develop other solutions to overcome this problem like overbooking to make up the costs of the clinics.[Lacy (2005)] But seems that its better finding the ways to overcome/prevent the issue of no show instead of implementing new methods as Overbooking indeed help to reduce the costs and increases efficiency but on the other hand it has a negative impact on the experience of the patient as it increases their waiting time and to avoid the longer wait-times people tend not showing up for their appointments.

ML-TECHNIQUES

"Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics" [Tom M. Mitchell]. Computer Science is all about creating new devices or techniques to solve issues. Statistics refers to analysis of the data with certain assumptions to consider. Whereas Machine Learning is a large chunk of both which would help us to train the machines to learn/perform on their own for solving the problems in more efficient way if possible than humans. It also integrates different methods and algorithms to understand better way of capturing data, understanding and performing required tasks." Machine Learning focuses on the question of how to get computers to program themselves (from experience plus some initial structure)" [Tom M. Mitchell].

Machine Learning mainly relates to various learning problems encountered after receiving the data like estimating, Predicting, Identifying the impact of risk etc. This plays a major role as every scenario has a different way of approach. Among all these learning problems, there are three main common types of Machine Learning methods. They are:

1. Supervised Learning: When you are training a data with certain features in an algorithm to estimate or predict the required output. In this the machine tries to understand the relationship between the various features with their names(labels).
2. Un-supervised Learning: The main aim in this type of learning is to find the clusters or patterns from the features as we have no measurements for the output/outcome.
3. Semi-supervised Learning: This type of learning is a combination of both supervised and un-supervised learning in analysing the labelled and unlabelled data to build a better learning model.

In this paper, we have a Supervised learning problem with a categorical outcome variable ("no-show" being "YES"/"NO") where "Yes" denotes that the patient did not show up for the appointment and "No" denotes that the person showed up for the appointment.

Feature Selection Method:

Genetic Algorithm: Genetic algorithm is an optimization function used to enhance the performance of the predictive models. This algorithm was evolved by the process of natural evolution. It is based on the theory of Darwin that states "Survival of the Fittest". As the population contains the attributes or independent variables out of which the best features would be selected for prediction.

The Fig (1) explains the complete process of selecting these features. The individuals are denoted as "0" and "1". The whole process is like the process of reproduction where the fittest individuals are selected for reproducing an off-spring by mutation and crossover so that the final off-spring would have the features of both the parents.

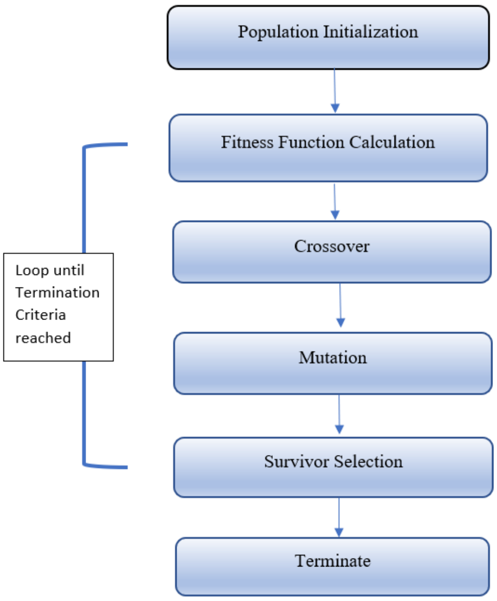

Population Initialization
Fitness Function Calculation
Crossover
Mutation
Survivor Selection
Terminate
Loop until Termination Criteria reached

Fig 1: Process of functionality of Genetic Algorithm

The following are the various models of prediction used in this paper:

Logistic regression:

Logistic regression is another technique of machine learning used mainly for predicting binary dependent attributes. It cannot be modelled directly by linear regression, as the response variable is discrete. It is a powerful tool that predicts the odds of its occurrence, it estimates the probability to belong to one category using a regression on independent variables. Hence, it is important to select the essential inputs and defining their relationship to the response variable.

Logistic regression model can be represented as:

$$Y(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where, Y= Dependent variable,

$X_1, X_2 \ldots . X_k$= Independent variables

$\beta_1, \beta_2 \ldots \beta_k$= regression co-efficient

P = probability of observation being 1 (No-Show=Yes)

β_0= Y intercept

ε = Error

Decision Tree:

The decision tree is one of the classic algorithms in predictive analysis to solve the binary classification issues. As the name suggests this method separates the observation in various branches to build a tree for improvement of prediction accuracy. Various mathematical algorithms (Eg: Chi-Squared test, Gini Index, information gain) are used to find a variable that splits the input observation into various sub-groups. These steps are repeated until the whole tree is built.

This advantage makes then widely usable in medicine (S.Dreiseitl, 2002). In this paper we used Gini index criterion for splitting. Gini Index is a measure for purity (or Impurity) which is used by the CART algorithm.

$$Gini_{index} = \sum_{i=1}^{n} w_i \, Gini_i$$

$$Gini_{impurity}(P) = 1 - \sum_{i=1}^{n} p_i^2$$

Here, Gini Index is defined as the waited sum of the Gini Impurity for the different sub-sets after a split. The feature with the lowest Gini Index is then used as the next splitting feature. When the tree grows the algorithm calculates selected features to find out which one will produce the best split. In case of numeric features, the split is always a binary split.

Artificial Neural Network (Multi-layered Perceptron):

Artificial Neural Networks is one of the complex statistical models that could be a better option than simple models like logistic Regression especially for this issue of predicting the factors impacting the no-show of patients in a healthcare environment. Feed-forward Network has been considered as first and simple Artificial Neural Network model. These networks have lately been very popular in resolving the real word problems especially in the health care sector. “Perceptron” is considered as an algorithm for the supervised learning that has binary classifiers. The fast-operational ability of MLP and better learning capacity in a smaller dataset was the reason to implement this model. Fig(2) represents the structure of an Artificial Neural Network with one hidden layer with various nodes that help in navigating the information from input layer to the Output layer with the required outcomes.

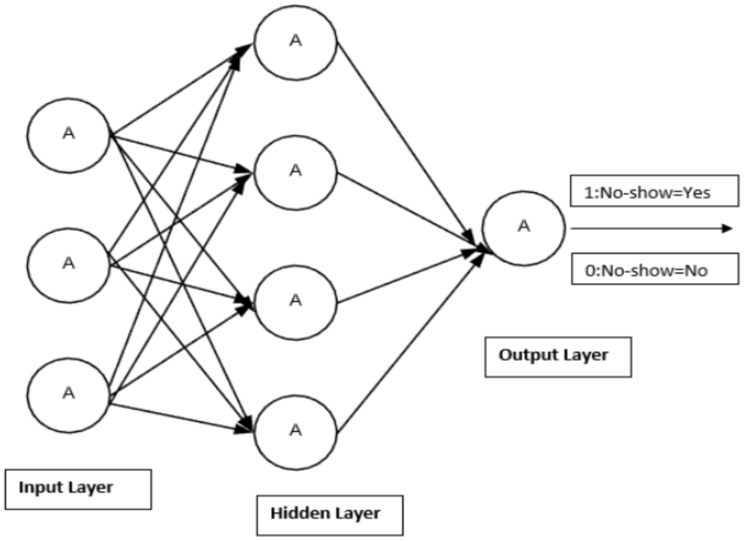


Fig 2: Structure of Artificial Neural Network with 1 hidden layer.

Random forest:

Ensemble Methods:

These are the methods that combine several individual models to get a better optimal predictive model or decrease variance or bias. These methods are categorised in two categories:

a. Sequential Ensemble Methods**:** In these methods the individual models are boosted by assigning weights to the previous mislabelled observations with higher weights.

b. Parallel Ensemble Methods**:** In these methods the individual models are combined, and trials are made to reduce the error rates by calculating the averages.

Random Forest is one of the Parallel Ensemble methods. It is considered as the supervised algorithm technique for both regression as well as Classification Problems. Random Forest builds an ensemble of trees to improve upon the limited robustness and suboptimal performance of decision trees (Dudoit. S. Fridlyand). Fig(3) represents the various tree structures in a random forest model.The main approach of these methods is to put together all the weak learners to form a strong learner. During the learning process in our research, "No-show" being the target column we selected all the other variables to use learn the model. "Gini Index" was selected as the split criterion.

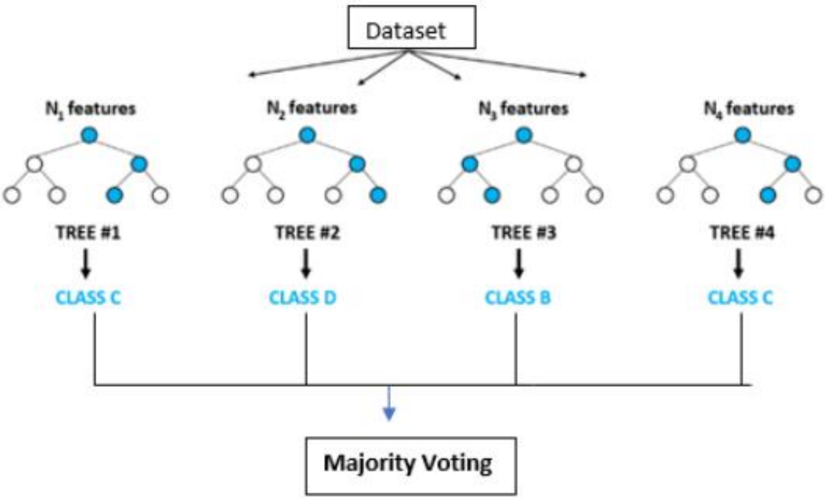


Fig 3: Tree Structures of Random Forest Model.

Bayesian Belief:

Bayesian Belief is a model that graphically represents the probabilities of the conditional dependencies of the attributes selecting randomly through a Directed Acyclic Graph (DAG). We decided to experiment this model to understand the casual relationships between the variables we selected and the target variable("No-show"). Also, these relationships can be helpful if the construction of other problems, and to infer techniques that could be applied to the results. It can be represented as:

$$P(X/Y)$$

If the variables are **Independent**, then

$$P(X/Y) = P(X)$$

If the variable is **Dependent**, then

$$P(X/Y) = \frac{P(X, Y)}{P(Y)}$$

Thus, the probabilities in Bayesian belief network can be calculated by:

$$\underbrace{P(X_1, \dots, X_N)}_{\text{Complete Distribution}} = \underbrace{\prod_{i=1}^{N} P(X_i|Parent(X_i))}_{\text{The graph structured estimation.}}$$

METHODOLOGY

This study examines the data from Kaggle's medical No-Show dataset Fig (4) represents the methodology used in this paper. We chose SAS jmp tool and R studio for the purpose of data-cleaning followed by feature engineering. Later, the dataset is partioned by 10-Fold Cross Validation for the purpose of generalization of model's performance. To tackle the imbalanced dataset, we used Random under Sampling (RUS) method and 5 different of models were experimented. On the other hand, for the purpose of comparison we also used Genetic Algorithm as Feature Selection method. All the models were executed in KNIME analytical tool. Accuracy, Sensitivity, Specificity, AUC were considered for evaluation of performance of the models. The overall results could be communicated and utilized for decision making.

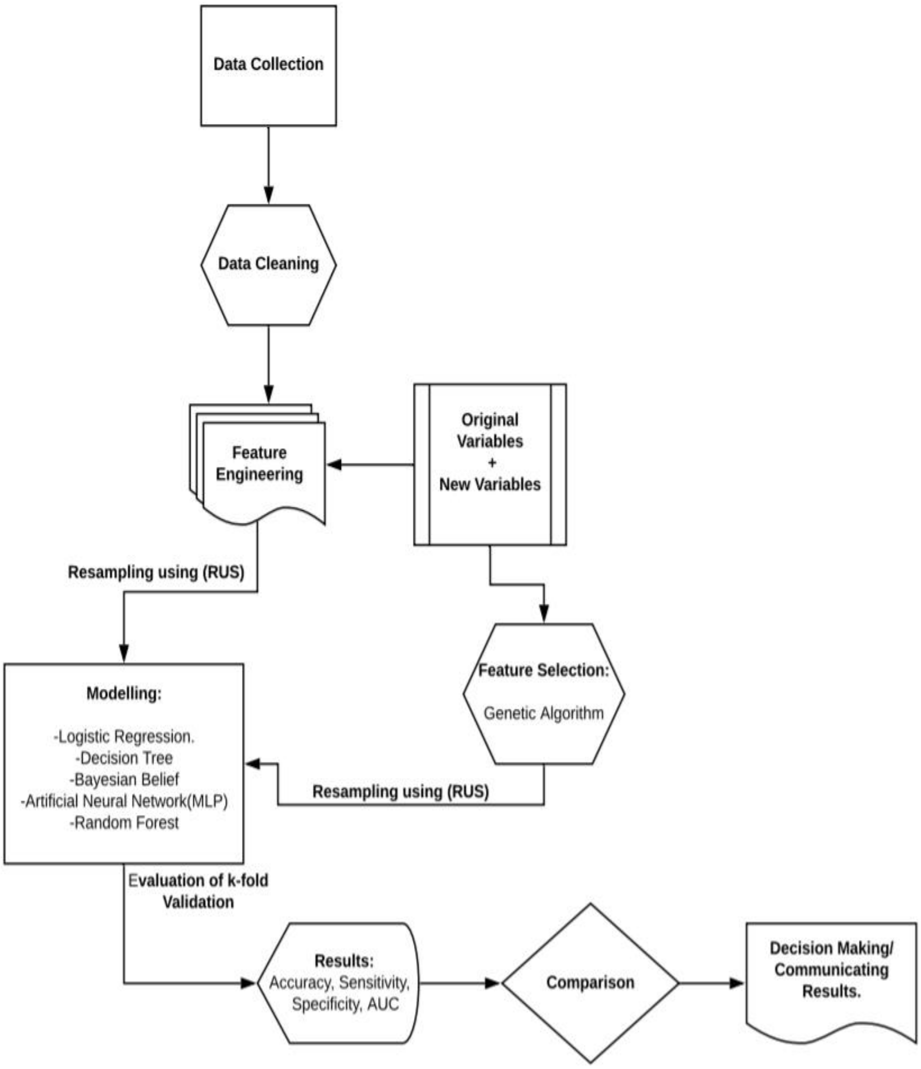

Data Collection
Data Cleaning
Feature Engineering
Original Variables + New Variables
Resampling using (RUS)
Feature Selection: Genetic Algorithm
Modelling: -Logistic Regression. -Decision Tree -Bayesian Belief -Artificial Neural Network(MLP) -Random Forest
Resampling using (RUS)
Evaluation of k-fold Validation
Results: Accuracy, Sensitivity, Specificity, AUC
Comparison
Decision Making/ Communicating Results.

Fig 4: Graphical representation of methodology used in this study.

DATA EXPLANATION AND RESEARCH

The dataset used for this analysis is from Kaggle by Joni Hoppen which is extracted from the Brazilian Public Health System known as Single Health System (SUS). This dataset contains 110,527 medical appointments stretched across a period of 2 months whereas the Scheduled dates are stretched across period of 6 months. It includes Patient level information including time of the appointment, when the appointment was scheduled, location of the clinic, whether SMS reminder was sent/or not, the outcome variable denoting if the patient showed up or not for the appointment.

The given below are the variables in the initial dataset with their descriptions:

Dataset attributes(categories):

PatientId: Unique Identification ID of the patient.

AppointmentID: Identification ID of each appointment booked by Patients.

Gender: Gender of Patient (Male or Female)

ScheduledDay: This is the day when patient registered/Confirmed the appointment.

AppointmentDay: This is the date of actual appointment when patient visits the clinic.

Age: Age of the patient.

Neighbourhood: This is the location where the health units are located.

Scholarship: Binary column where "1" denote the patient ID's who receives the subsidy (Bolsa Familia Program) & "0" denote the patient ID's that don't receive the subsidy (Bolsa Familia Program).

Hypertension: Binary column where "1" denote Patient ID's having this condition and "0" denote the Patient ID's that don't have this condition.

Diabetes: Binary Column where "1" denote Patient ID's having this condition and "0" denote the Patient ID's that don't have this condition.

Alcoholism: Binary Column where "1" denote Patient ID's consuming Alcohol and "0" denote the Patient ID's who don't consume Alcohol.

Handcap: This column ranges from 0 to 4 that denote sum of Conditions each Patient ID have (Such as deaf, blind etc)

SMS_received: Binary Column where "1" denote SMS was received by a patient and "0" denote that the Patient didn't receive SMS.

No-show (Target Column): Binary Column where "1" denote that the patient didn't show up for the appointment and "0" denote that the patient showed up for the appointment.

The initial phase of data-cleaning began with re-naming the mis-spelled variables. Most of the data cleaning was performed in R Studio and SAS JMP. A strange but real fact being no-missing values were found in the dataset. Given below are few of the interesting distributions of the variables for better understanding: (Before data- cleaning)

Fig 5: Distributions (Before Data-cleaning)

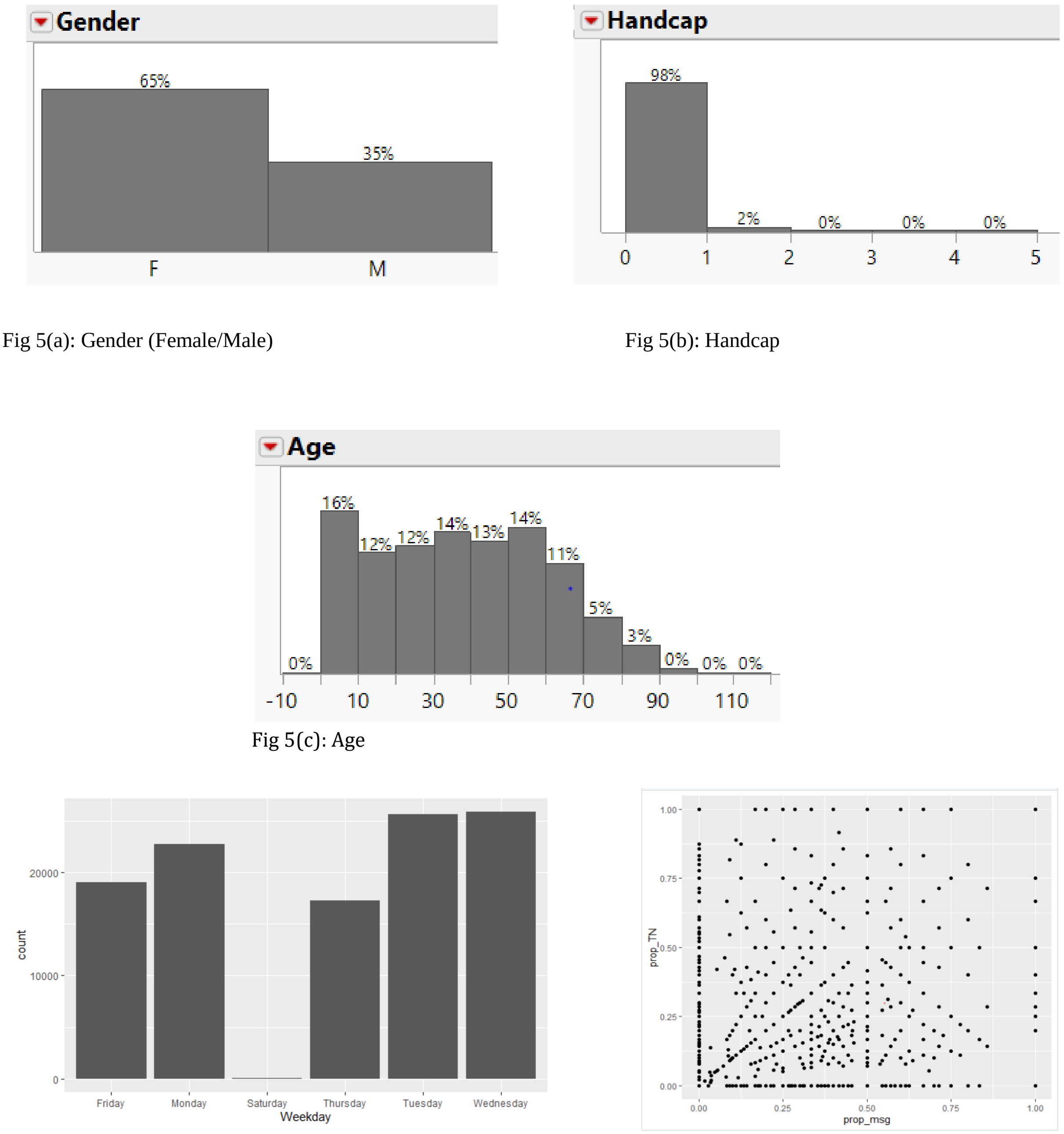


Fig 5(a): Gender (Female/Male)

Fig 5(b): Handcap

Fig 5(c): Age

Fig 5(d): Weekdays of Appointment day

Fig 5(e): Scatter plot of SMS_received v/s No-shows

The above Fig 5(a) shows that the proportion of Females (65%) is higher than the proportion of males (35%). The high ratio of the female patients could denote that the appointments are related to the pregnancy. The mean age of the females is 38years and males is 33years. Later this Column was coded as Female=”1” and Male=”0”. Fig 5(b) denotes one of the nuances in the dataset. The column “Handcap” was assumed to be a binary but in fact it is a categorical variable ranging from 0 to 5 denoting the sum of conditions each patient could have. For example, if a patient is deaf it is considered as “1”, if a patient is deaf, blind this column would denote as “2” and so on. Less than #% of the dataset have the patients with more than 2 conditions. Fig 5(c) denotes the column of age of the patients. The graph on the x-axis denotes the ages of the patients with 10 being the bin size. One of the age being negative (-10) is later removed from the dataset considering it to be wrongly entered. The age 0 could be considered as newly born kids and thus the higher number of visits for younger age group is considerable. Later a new variable is created from this column i.e, Age_Grp where these ages are categorised into three kinds [1= Kids,2=Adults,3=Old]. Kids ranged from 0 to 15years, Adults ranged from 16years to 45years and Old being greater than 45years.

Fig5(d) represents the weekdays of the appointment date when the patients visited the clinics, no patients were recorded on Sunday. The proportion of patient visits seemed increasing at the very beginning of the week i.e., Monday to Friday and, the proportion gradually decreases on the weekends. It could be a scenario where the clinics might only accept emergency cases on the weekends. Fig5(e) represents a Scatter plot graph between the proportion of SMS_received (prop_msg) by each patient and the proportion of No-Shows (prop_TN). The correlation between the Prop_msg and Prop_TN was 0.128 which was expected to be higher but looks like this isn’t the main factor impacting the no show of patients. If the graph is divided into 4 quadrants (Q1, Q2, Q3, Q4), Q3 &Q4 seem to accumulate higher number of points i.e., they have considerate number of No-Shows when the propitiate number of SMS_received is between 0 to 0.5. But on the other hand, the data points are scattered all over the 4 quadrants, this could also be impacted by certain outliers which are the No-Shows that can be seen when prop_msg=1

After exploring the data, we thought generation of some new variables from the existing ones could be helpful in the modelling for improving the accuracy rates. The following are the new variables created in the phase of feature extraction:

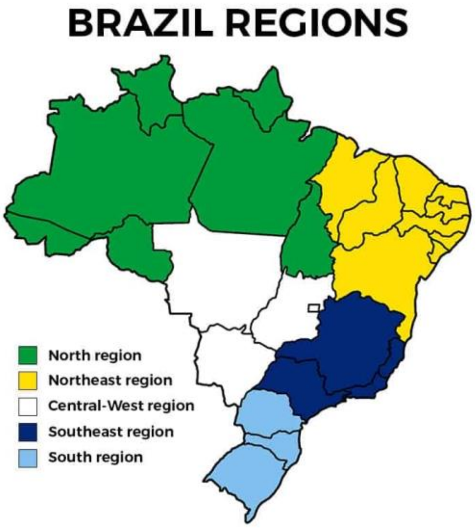
- wait_time
- Age_grp
- region
- prior_appointments
- prior_noshows
- pa_pn

wait_time: This column denotes the number of days between the scheduled day and the appointment day of the patients visiting the clinic.

wait_time = Appointment day – Schedule day

Age_grp: This column is created by categorizing the ages of the patients into 3 levels namely Kids=1 (0-15yrs), Adult=2 (16yrs- 45yrs), Old=3 (>45yrs).

region: There were 81 unique neighbourhoods where the clinics were located. Most of the neighbourhoods belong to city of Vitoria in Brazil, but it also included few other neighbouring cities. To reduce the complexity for modelling we grouped the neighbourhoods into 5 regions namely:



- North Region
- North-East Region
- Central-west Region
- South-East Region
- South Region

Fig 5: Region of Brazil

Prior_appointments: This column denotes the count of appointments excluding the last appointment as we are trying to predict the number of No-Shows for last appointment of each patient.

Prior_appointments= Total Appointments (TA) - 1

prior_noshows: This column denotes the number of No-Shows by each patient excluding the status last appointment i.e., as follows:

Prior_noshows = Total noshows – last noshow

Pa_pn: This is an interaction term between the number of prior appointments and the number of prior no shows. Pa_pn was the new variable created when we tested these two columns by calculating the product and found that it was significant. The formula for calculating this is as follows:

Pa_pn = Prior_appointments – Prior_noshows

Given below are some of the distributions of the new variables created which would be included for the further modelling.

Fig 6: Distributions

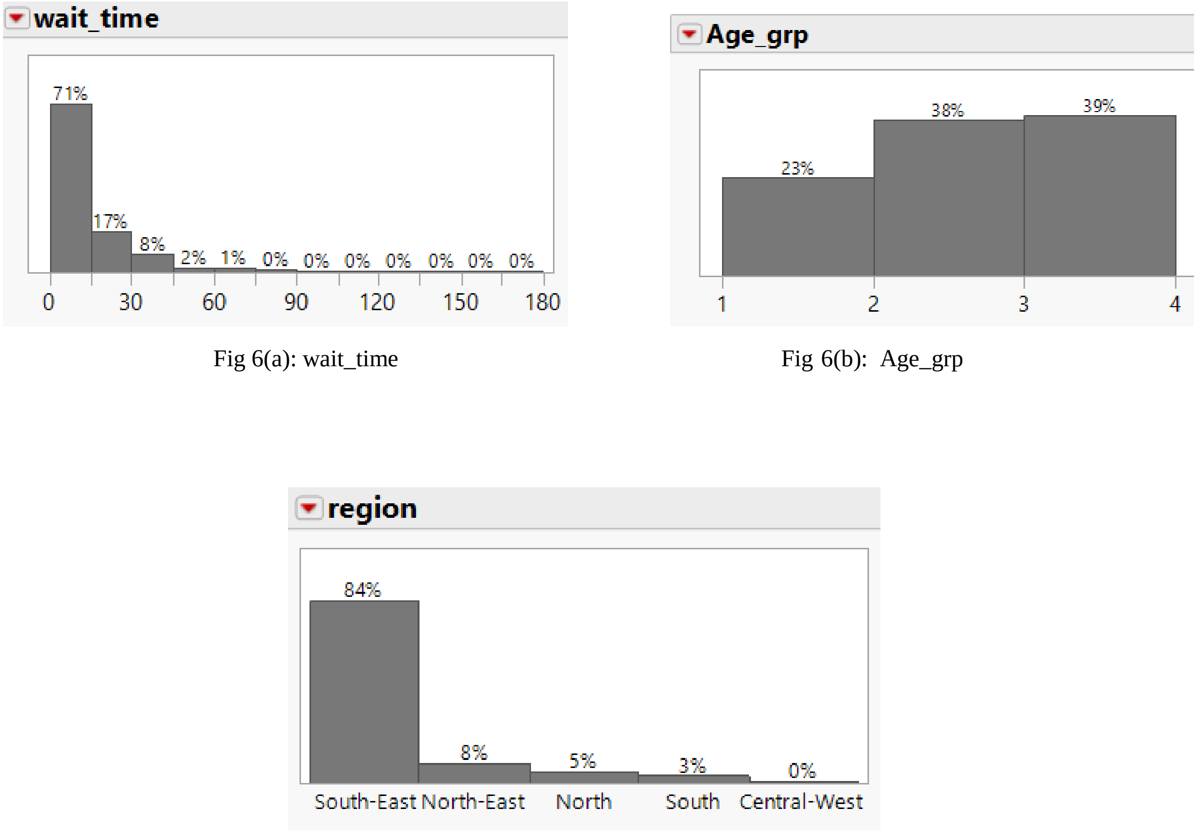


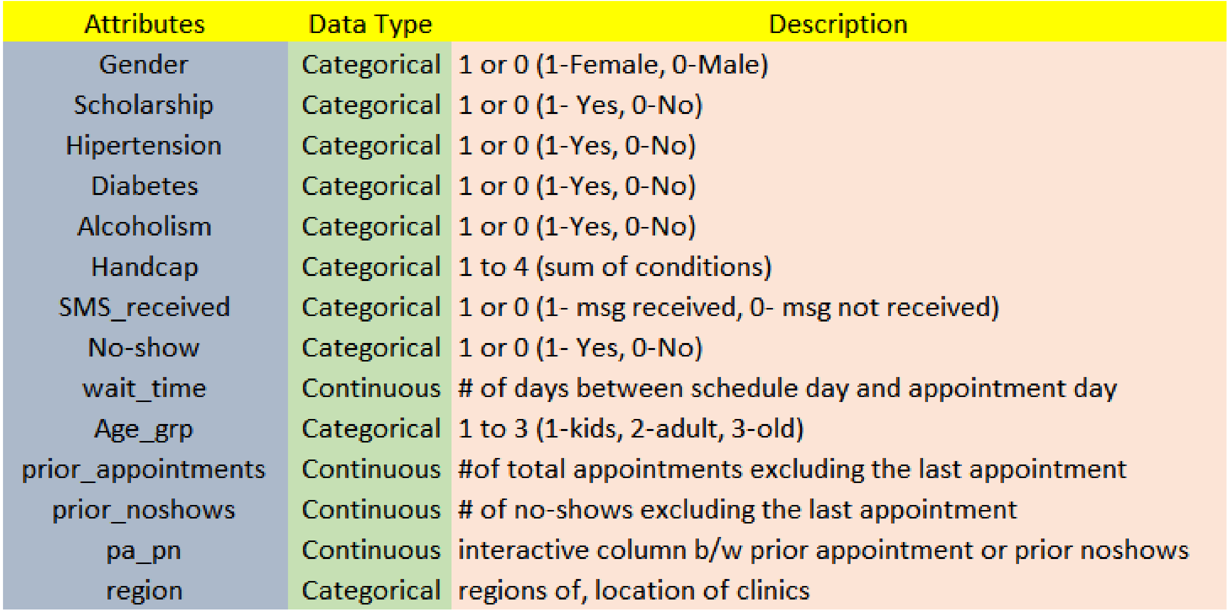
Fig 6(a): wait_time

Fig 6(b): Age_grp

Fig 6(c): region

The above Fig 6 are few of the distributions of the new variables we created with the existing variables. Fig 6(a) denotes the wait_time between the scheduled day and the appointment day which patient attends the clinic. It is also known as lead time. The figure shows that 71% of the patients were having a wait_time that are less than 15 days. The patients denoting “0” are the ones who take the appointments for the same day. In the Fig 6(b) the bar from 1 to 2 denotes the category “1” (Kids) which are 23% in the dataset. The highest proportion from 3 to 4 being the “3” (Old) with 39%. It is considerable as we could say that older people tend to have higher number of clinic visits when compared to adults as well as Kids. Fig 6(c) represents the grouping of 81 unique neighbourhoods that were grouped as 5 regions. As mentioned before, the figure represents that most of the neighbourhoods belonged to the city of Vitória that lies in the South-East region of Brazil consisting almost 84% of the total dataset. But 15% of the neighbourhoods were from other neighbouring places in brazil.

The following mentioned in Fig 7 are the variables (Feature Selection + Feature Extraction) that were used for the modelling with their datatypes and description.



Attributes	Data Type	Description
Gender	Categorical	1 or 0 (1-Female, 0-Male)
Scholarship	Categorical	1 or 0 (1- Yes, 0-No)
Hipertension	Categorical	1 or 0 (1-Yes, 0-No)
Diabetes	Categorical	1 or 0 (1-Yes, 0-No)
Alcoholism	Categorical	1 or 0 (1-Yes, 0-No)
Handcap	Categorical	1 to 4 (sum of conditions)
SMS_received	Categorical	1 or 0 (1- msg received, 0- msg not received)
No-show	Categorical	1 or 0 (1- Yes, 0-No)
wait_time	Continuous	# of days between schedule day and appointment day
Age_grp	Categorical	1 to 3 (1-kids, 2-adult, 3-old)
prior_appointments	Continuous	#of total appointments excluding the last appointment
prior_noshows	Continuous	# of no-shows excluding the last appointment
pa_pn	Continuous	interactive column b/w prior appointment or prior noshows
region	Categorical	regions of, location of clinics

Fig 7: Description of all the variables used further for modelling.

K-fold cross-Validation:

Before experimenting the models, the Data was split into ‘K’(K=10) number of training/test sets by stratified sampling. “K” value being 10 is considered by various experiments that the model result would denote moderate variance and less bias. This procedure is considered to give more impact to predict the competency of machine learning models on an unseen data when compared to simple split method where the data is just split into two parts being training and test set. Fig 8 represents the 10-fold cross validation. The following procedure is followed in k-fold cross validation:

- The data is rearranged/jumbled randomly.
- Then it is divided into k parts(k=10)
- For each time, one part is set aside as test set and (k-1) parts are considered as training set.
- For each time model is fit on the training set and evaluated on the test set.
- This process continues for “K” times and the scores are retained discarding the model
- These are summarized using the model evaluation scores.

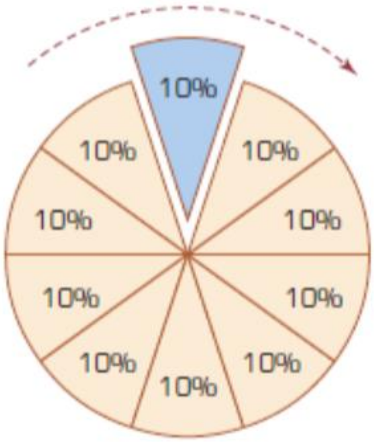


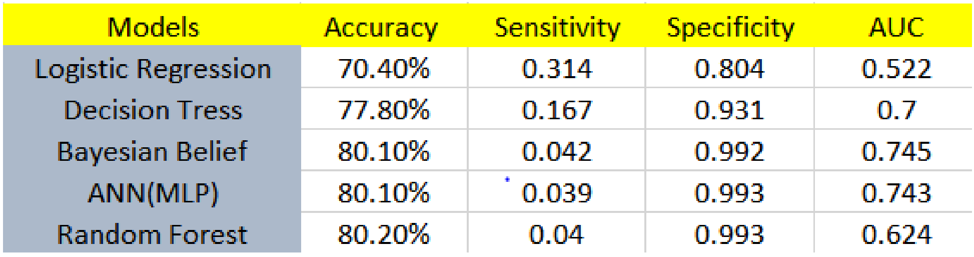
Fig 8: Graphical Representation of 10-fold cross-validation

RESULTS:

Experiment-1:

Modelling and comparison of 5 different models using all the features mentioned in Fig:

Among five different models, 62299 unique patients were there in the final dataset where three models namely Bayesian Belief(AUC=0.745), Artificial Neural Networks(MLP)(AUC=0.743),Random Forest(AUC=0.624) were considered to perform well in predicting the last appointment's no show by training the dataset with the data of prior appointments with accuracies around 80%. 49928 patients were correctly classified, and the Fig 9 below shows that these models had higher specificity levels which is the true negative rate I.e, Predicting the # of patients that showed up were correctly classified, but sensitivity rates seem very low.



Models	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	70.40%	0.314	0.804	0.522
Decision Tress	77.80%	0.167	0.931	0.7
Bayesian Belief	80.10%	0.042	0.992	0.745
ANN(MLP)	80.10%	0.039	0.993	0.743
Random Forest	80.20%	0.04	0.993	0.624

Fig 9 : Model Results using all the features.

The above Fig 9 showed better accuracy levels, but we observed that there was a severe skew in the distribution of target column "No-show". So, we resampled the dataset after splitting and before running the model. We used Random Under sampling (RUS) to handle the imbalanced dataset. Random under sampling is a simplest strategy to randomly removing the examples from the majority class in the training set. Fig (10) represents the target column before and after under sampling the majority class of the target variable.

Fig 10: Distributions of target column.

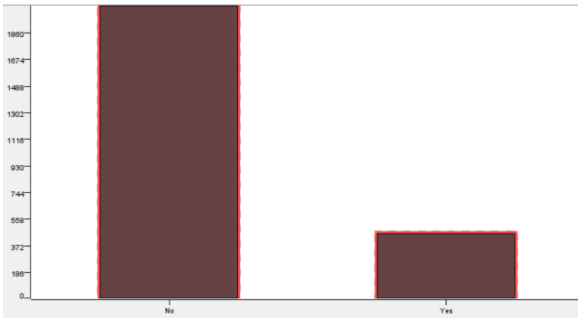


Fig 10(a): Before re-sampling using RUS

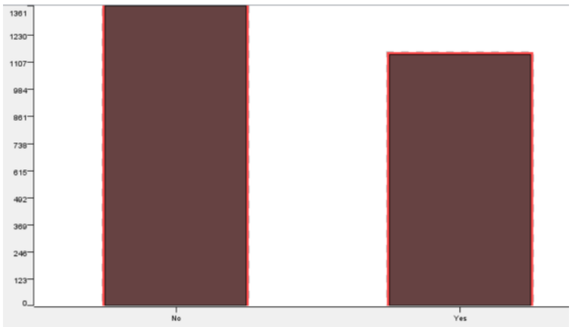
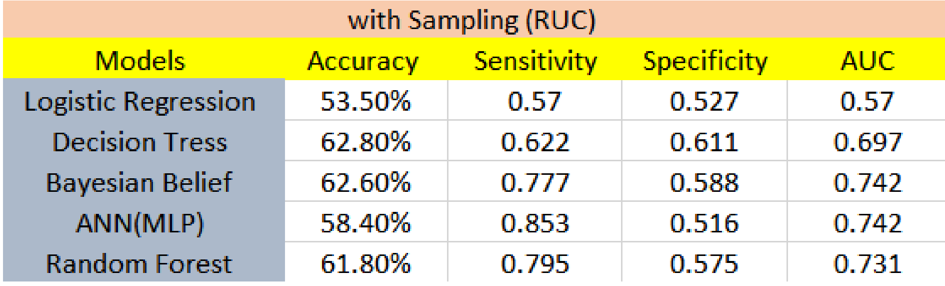


Fig 10(b): After re-sampling using RUS.

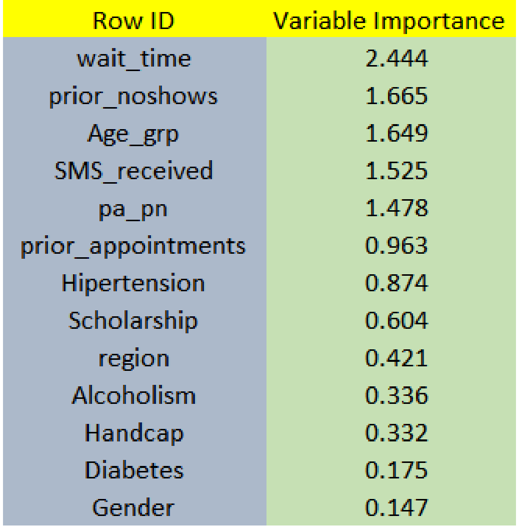
Experiment-2:

Given below in the Fig 11 are the results of all the mentioned models after re-sampling the dataset using Random Under sampling:



with Sampling (RUC)				
Models	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	53.50%	0.57	0.527	0.57
Decision Tress	62.80%	0.622	0.611	0.697
Bayesian Belief	62.60%	0.777	0.588	0.742
ANN(MLP)	58.40%	0.853	0.516	0.742
Random Forest	61.80%	0.795	0.575	0.731

Fig 11: Model Results using all the features after Sampling using RUS



Row ID	Variable Importance
wait_time	2.444
prior_noshows	1.665
Age_grp	1.649
SMS_received	1.525
pa_pn	1.478
prior_appointments	0.963
Hipertension	0.874
Scholarship	0.604
region	0.421
Alcoholism	0.336
Handcap	0.332
Diabetes	0.175
Gender	0.147

Fig 12: Representation of Variable Importance (Random Forest)

Above Fig 12 shows that after sampling the dataset still Bayesian Belief,Artificial Neural Network and Random Forest with AUC=0.742,AUC=0.742,AUC=0.731 respectively perform better than the other models though their accuracy levels reduced from 80.10% to 62.60% and 58.40% respectively.It is also observed that there is slight improvement in the sensitivity rates i.e, Proportion of correctly classifying the patients who did not show up at the clinic.The Fig represents the variable Importance of the attributes. Wait_time being the most significnt variable with 2.4444.These values would be important to understand and analyse the factors that highly impact the no show of patients to the clinics. The Fig 13 represents the AUC curve of Bayesian Belief and Artificial Neural Network (AUC = 0.742), Which were the highest among all the other models.

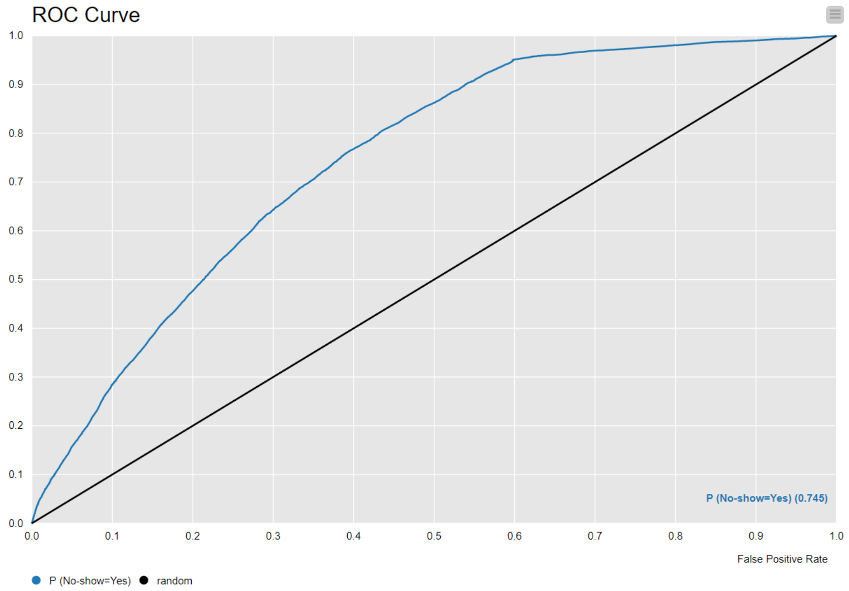


Fig 13: ROC curve of Bayesian belief and Artificial Neural Network.

The Fig 14 below is the tree augmented Network which represents the interrelations between the features that highly impact the target variable(“No-show”= Yes). The diagram represents two categories of variabls, namely

- Parent Features(Nodes)
- Child Features(Nodes)

For example, “prior-appointments” node seems to have direct relation with the target variable(No-show) and thus considered as Parent Node and “wait_time”, “pa_pn” seem to be related to “No-show” through the parent node “prior_appointments”, thus it is considered as Child Node.Similarly this tree augmented Network could be useful to understand the relationships between the attributes.

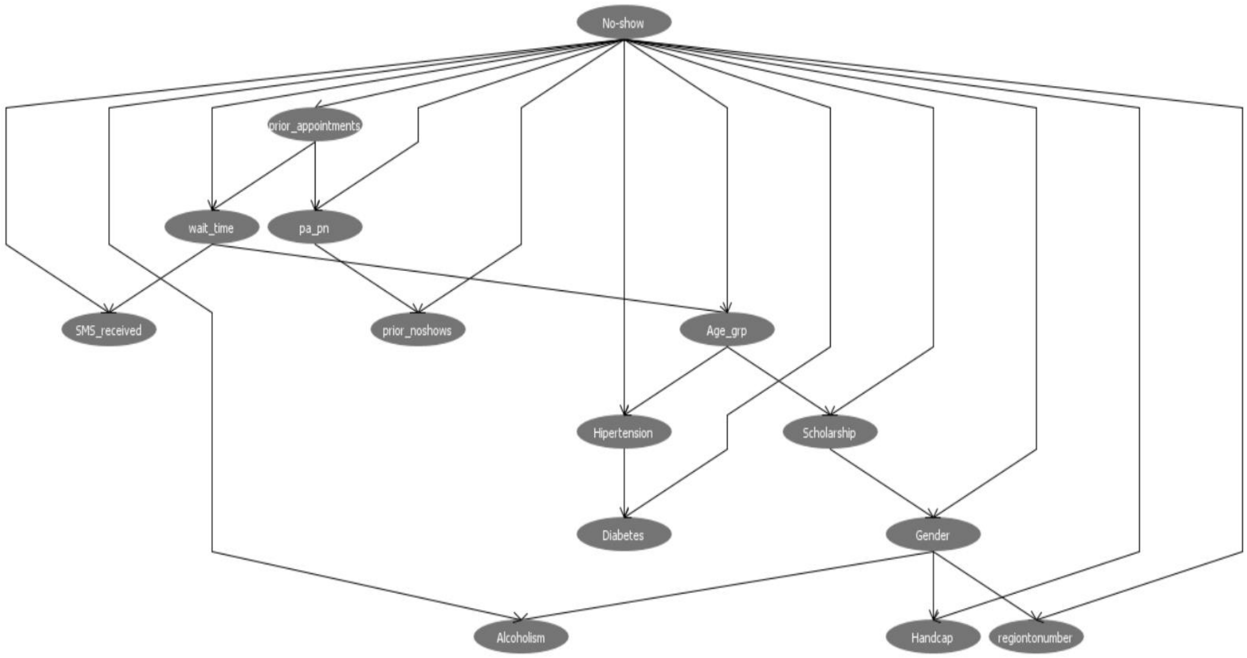


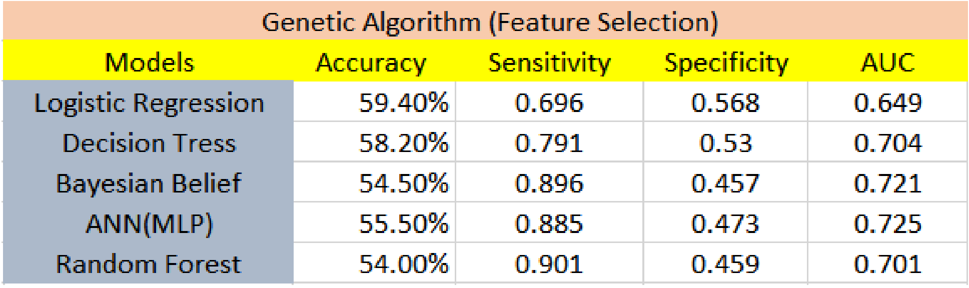
Fig 14: Augmented Tree Network (Bayesian Belief Network)

Experiment -3:

Another experiment was performed to see it there would be possibility in improving the AUC values by using a feature selection method. In this case we applied the genetic algorithm with parameters of population size=100, # of generations=15. The range of variables were set from 5 to 9. From the subsets of the results, the variables chosen for the modelling were as follows:

- Scholarship
- Hipertension
- Alcoholism
- Handicap
- SMS_received
- Wait_time
- Prior_appointment
- Region
- No-show (Target Column)

Results of mentioned models using Feature Selection method (Genetic Algorithm):



Genetic Algorithm (Feature Selection)

Models	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	59.40%	0.696	0.568	0.649
Decision Tress	58.20%	0.791	0.53	0.704
Bayesian Belief	54.50%	0.896	0.457	0.721
ANN(MLP)	55.50%	0.885	0.473	0.725
Random Forest	54.00%	0.901	0.459	0.701

Fig 15: Model Results using Feature Selection method(Genetic Algorithm)

Above Fig 15 shows after using feature selection, the AUC values gradually decreased for almost all the models but the sensitivity highly increased for almost all the models. The highest being Random Forest from 0.795 to 0.901 i.e., the proportion of patients not showing up at the clinic were correctly classified. In this dataset scenario feature selection did not show better result when compared to models using all the features for modelling. One of the reasons could be the low number of total features being used in the model. The idea of feature selection could be helpful when you have higher # of features in the dataset.

DISCUSSION

This study highlighted the significant factors of no-show of patients of Healthcare units in Brazil by predicting the last appointment of the patient by training various models with the prior appointments. It showed that the status of their prior visits has a higher impact on their next visit and the patients scheduling their appointments on the same day were the most important predictors of patients not showing up in the clinics. The results showed that the best model being Bayesian Belief AUC= 0.742 were able to predict around 62.6% of accuracy and were able to classify around 38999 patients out of 62299 Patients.

Like the previous research, our results also stated wait_time and prior_appointments being the most significant factors (Daggy J 2010). For the further research few other factors could also be taken into consideration like the weather, demographic factors. Thus, increasing the number of variables would surely help in improving the better results that would further be helpful for the clinics to analyse and get a overview to implement the steps to reduce the level of no-shows. The study also showed that the column of SMS_received didn’t have a huge impact in impacting the no-show, this could help the clinics to understand and give a thought of reducing their clinics cost by not spending much on sending additional reminders to the patient. As this study shows that same day appointments had a major impact on the No-Show of patients thus, Health Care units can imply considering only significant or emergency related patients for scheduling an appointment on same day. Weather, Seasons could also be a significant topic to be considered for the purpose of further research.

CONCLUSION

This study developed a comparative research of predicting various models in predicting the last appointment no-show of patients by training the data of their previous appointments from the data of Health care units in Brazil. These models could be used by not just the healthcare centres but also any sector that follow the system of scheduling. This study is helpful not just to predict the proportion of no-show but also to analyse their data in understanding and predicting the significant factors impacting the patients not visiting the clinics. This would be helpful to them to identify the lack of area in their field and to give scope of improvement that could help them in gaining huge profits and in Cost reduction techniques. Additional patient information like income levels, details regarding educational background, patient's location would be helpful in improving the accuracy levels of the model. The clinics could also implement a process of feedback forms from the patients with few details about their last appointment and the reason(if they missed it)that would help us understand from the perspective of patient and would help us understand analyse in better ways to reduce the no-shows.

ACKNOWLEDGMENT

We would like to thank Professor Eyyub Y. Kibis, Ph.D. for his expert advice and encouragement through out this project.

REFERENCES

[1] GeorgeA,Rubin G. Non-attendance in general practice: a systematic review and its implications for access to primary health care. Fam Pract 2003; 20: 178-184. Doi: 10.1093/fampra/20.22178 12651793

[2] Alessandra Trindade Machado, Marcos Azeredo Furquim Werneck, Simone Dutra Lucas, Mauro Henrique Nogueira Guimaraes Abreu : Who did not appear? First dental visit absences in secondary care in a major o Brazilian city: a cross-sectional study.

[3] Daggy J, Lawley M, Willis D, et al: Using no-show modeling to improve clinic performance. Health Informatics J 2010; 16(4): 246–59.

[4] Tom M. Mitchell: The Discipline of Machine Learning.

[5] Lacy 2005:Medical and Mathematical Authorship in Ancient Greece.

[6] Henry Lenzi,Angela Jornada Ben, Airton Tetelbom Stein: Development and validation of a patientno-show predictive model at a primary care setting in Southern Brazil.

[7] Virginia Mato Abad, Isabel Jimenez, Rafael Garcia-Vazquez, Santiago Rodriguez: Using Artificial Neural Networks for Identifying Patients with Mild Cognitive Impairment Associated with Depression Using Neuropsychological Test Features - Scientific Figure on ResearchGate.

[8] Dudoit. S. Fridlyand: Comparison of discrimination methods for the classification of tumours using Gene expression data.

[9] S.Dreiseitl, L.Ohno-Machado: Logistic regression and artificial neural network classification models: a methodology review, Journal of Bio-medical informatics (2002)

[10] Mohammadi, A Turkcan, T Toscos, A Miller, K Kunjan. Assessing and Simulating Processes in Community Health Centers. -AMIA (2015)

[11] Iman Mohammadi, Huanmei Wu, Ayten, Turkcan, Tammy Toscos, Bradley N. Doebbeling. Data analytics and modeling for appointment No-Show in community health centers.

[12] Leila F. Dantas, Julia L. Fleck, Fernando L. Cyrino Oliveira, Silvio Hamacher. No- show in appointment scheduling- a sysyematic literture review.

[13] Ronald C. Samuels, MD, MPH, Valerie L. Ward, MD, MPH, Patrice Melvin, MPH, Micharl Macht-Greenberg. Missing appointments: Factors contributing to high No-Show rates in an urban pediatrics primary care clinic. (2015)

[14] Philip J. Tuso MD FACP, ken Murtishaw RN MA DHE, Wadie MD. Way to reduce patients No-Show rate, Decrease add-one to primary care schedules, and improve patient satisfaction.

[15] Bhagwan Satiani MD, MBA, FACHE, Susan Miller RVT, Darshan Patel. No-Show rates in the vascular laboratory: Analysis and possible solutions.

[16] Thomas Vikander, MD, Kris Parnicky, MD, Raymond Demers, MD, MPH, Kenneth Frisof, MD, Paul Demers, and Nathan Chase. New-Patients No-Shows in an Urban family practice center: Analysis and Intervention.

[17] Douglas L. Nguyen, MD, Ramona S. DJesus, MD, Mark L. Wieland, MD, MPH. Misses appointments in resident continuity clinic: patient characteristics and health care outcomes.

[18] Somayeh Anisi, Ehsan Zarei, Mahnaz Sabzi, Mohammad Chehrazi. Missed appointments: Factors contribution to patient No-Show in outpatients' hospitals clinics.

[19] Joanne daggy and Mark Lawley, Deanna willis, Debrs Thayer and Christopher Suelzer, Po-Ching DeLaurentis, Ayten Turkcan, Santanu Chakraborty and Laura Sands. Using No-Show modeling to improve clinic performance.

[20] Ashwin Mehra PHD MBA, Claire J. Joogendoorn PHD, Greg Haggerty PHD, Jessica Engeithaler BS, Stephen Gooden MHA, Michelle Joseph MSHA, Shannom Carroll DO, Peter A. Guiney DO. Reducing patient No-Shows: An initiative at an integrate care teaching health center.

[21] Caitlin E. Fiorillo MD, Allyson L. Hughes BS, Chen-I-Chen MS MPH, Philip M.Westgate PHD, Thomas J.Gal MD MPH, Matthew L.Bush MD, Brett T. comer MD. Factors associated with patient No-Show rates in an academic otolaryngology practice.

[22] Lee Goldman MD MPH, Ralph Freidin MD, E. Francis Cook MS. A multivariant approach to the prediction of No-Show behavior in a primary care center.

[23] Todd Molfenter. Reducing appointment No-Show going from theory to practice.

[24] Andrew S.Hwang BS, Steven J.Atlas MD MPH, Patrick Cronin MA, Jeffrey M, Ashburner MPH, Sachin J.Shah MD, Wei He MS & Clemens S.Hong MDMMPH. Appointment No-Show are an independent

[25] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Frichen Shen. Clinical information extraction applications: A literature review.

[26] Emma Kaplan-Lewis, Sanjat Percac-Lime. No-Show to primary care appointment: Why patients do not come.

[27] Stephan Dreiseitl, Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: A methodology review.

[28] Orlando Torres, Michael B. Rothberg, Jane Garb, Owolabi Ogunneye, Judepatricks Onyema, Thomas Higgins. Risk factors models to predict a missed clinic appointment in an urban, academic, and underserved setting.

[29] Y Hung, P Zuniga. Effective cancellation policy to reduce the negative impact of patient no-show.