

Analysis on No-Show Patient Appointments for Hospital X

Loh Yan Zoey; Mirania Aishwarya Agarwal; Nasrullah Bin Khairullah;

Singapore Management University

ABSTRACT

The healthcare industry in Singapore has always been keen on gathering insights on patients and their no-show appointments at their clinics. Hospitals are adopting a multi-pronged approach in reducing no-show appointments by improving the efficiency of appointment system and enabling patients to make changes of appointments more easily. The number of no-show appointments has an impact on the operational costs and clinic utilization. It also presents an opportunity cost for another patient who is unable to make use of the no-show appointment slot to get a consultation from a doctor or an allied health professional. In this study, we aim to identify significant factors that affect no-show appointments in a hospital. Taking references from past literature review, we will select and derive relevant variables to be used for modelling a possible solution for the problem being studied. We will develop models to predict the probability of no-shows for a hospital using both patient information and individual clinical appointment attendance records. We will then compare the different models and assess the results.

Based on our findings, we will end the report with set of implications and results for a hospital.

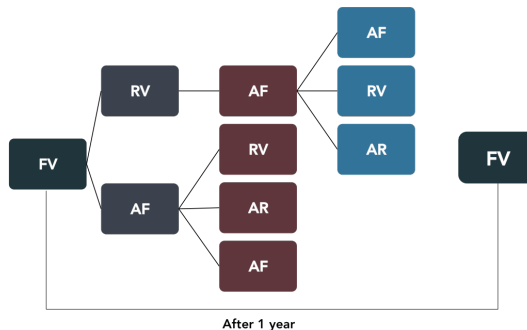
1.0 INTRODUCTION

With regards to the state of mental health disorders in children, there has been an increase of cases from 533 in 1980 to 3051 in 2010. A medical study (Woo, et al, 2007) has shown that one in eight children in Singapore has emotional disorders, and one in 20 has behavioural disorders, only 10% ever see a psychiatrist. Thus, it places an emphasis in understanding no-show appointments. Appointments are made for a reason. When patients default on their appointments, they miss the opportunity for a medical consultation and thus, place their health status at risk.

No-show appointment is defined as when a patient does not attend for a scheduled clinic appointment or cancels with such minimal lead time that the slot cannot be filled (Huang & Hanauer, 2014). The impact of no-show appointments includes disruption of efficient operations of the clinics, provider productivity, decreased access to care and depriving other patients of the opportunity to see a medical professional during no-show appointments.

1.1 PROJECT BACKGROUND

Hospital X is a pioneer tertiary hospital that provides a comprehensive range of medical and rehabilitative services for children, adolescents, adults and the elderly. Patients are usually referred to Hospital X by other medical institutions or they booked an appointment directly. Patients can be categorised according to their appointments with a doctor, an allied health professional or even both.



As seen in Figure 1, a patient's first appointment begins with a diagnosis by a doctor and subsequent appointments are made according to the patient's mental health status. If a patient does not have any appointment for a year, any subsequent appointment will have to be diagnosed by a doctor again (FV).

Our project sponsor is a medical consultant working for Hospital X. He specialises in tending to younger patients from the age of 18 years old and below. He hopes to tap into the under-utilised administrative data that is collected by the hospital daily.

According to our project sponsor, Hospital X experiences high no-show appointments rate of about 21% for first visits and 19% for review visits. Our project sponsor is keen on improving the access to care as missed appointments lead to longer appointment lead times, idle time and an overall reduced quality of care.

This paper seeks to explore the no-show patterns of the patients' appointments in Hospital X from 2015 to 2016.

1.2 PROJECT MOTIVATION

The project offers a unique opportunity to explore an unfamiliar domain (healthcare sector). We can learn much from our project sponsor as he is a champion for data analytics and has considerable experience.

People's behaviour, even when they follow expectations, is often varied and unique, and there is always a possibility of observing unexpected behaviours in the data. While we start off seeking to find patterns in no-show appointments, there is no telling what our analysis of the data may uncover about the situation and the problem, and that makes the project interesting.

It is very encouraging that our work has the potential to impact people's lives. The sponsor's concern for the community which he works in is motivating as well. That has led him to do more for the hospital, even beyond his normal duties, inspires us to study the data and see if we can find anything that could help the hospital to improve its processes.

1.3 OBJECTIVES

The objectives of the project would be of the following:

1. Business objective: To identify factors that relate to no-show appointments and predict patients' attendance rate in order to improve Hospital X's scheduling of appointments and utilisation of appointment slots.
2. Technical objective: To use data analytics tools and statistical methods to study the data and obtain insights that would facilitate the business objective.
 - To understand the data domains
 - To understand the workflow of scheduling a patient's consultation process
 - To identify the contributing factors that lead patients to defaulting appointments
 - To create a predictive model

1.4 PAPER OUTLINE

The rest of the paper is organized as follows. Section 2.0 reviews past literature papers that are related to our study. Section 3.0 elaborates our methodology as well as analytical methods. Section 4.0 showcases our exploratory data analysis as well as model results. Section 5.0 discusses the results, limitations and future implications to Hospital X. Section 6.0 will summarize our findings and offer future directions.

2.0 LITERATURE REVIEW

Ma, Seemanta, Wu and Ng (2014) developed logistic regression and recursive partitioning models, using SAP records to predict patients' no-show probabilities for each of the three clinics. The study included external information such as financial debt and reminder responses as predictor variables for no-show probability of patients. The results showed that there were some variations in the main predictor variables for no-show appointments among the three clinics.

Allaeddini, Yang, Reddy, Yu (2011) developed a hybrid probabilistic model that combines logistic regression as a population-based approach along with Bayesian inference as an individual-based approach for the no-show prediction model. The model included the effect of appointment characteristics such as number of previous appointments, appointment types and lead times in the next scheduled appointment. The study also highlighted that there are other types of disruption such as cancellation of appointments and patient lateness that may have an impact on the performance of the scheduling system.

Huang and Hanauer (2014) developed an evidence-based predictive model for no-show appointments and to improve overbooking approaches in outpatient settings to reduce the negative impact of no-shows. Factors like distance to the clinic, appointment characteristics, general demographic information and insurance information have been considered. One unique variable that this study has taken into account is the number of people in the household of the patient.

William, M.S.W and BCD (2001) provided explanations to deepen practitioners' understanding and management of no-show appointments. The study showed that no-show behaviour is positively correlated with lower income, lower socioeconomic status and lower age. Patients with more serious psychological difficulties are particularly taxed by long waiting times.

Michael et al. (2016) described patterns of no-show variation by patient age, gender, appointment age, and type of appointment request, using eight years' worth of individual-level records. A multifactor analysis of variance (ANOVA) was performed to characterize no-show and attendance rates and the impact of certain patient factors. One of the findings showed that the longer a patient has to wait for an appointment to be scheduled, the less likely is the patient to keep the first appointment.

A key distinction between our project and the literature review is that our project's appointments can be further broken down into consultation with a doctor or an allied health professional. The reference [Ma, Seemanta, Wu and Ng, 2014] is especially relevant and similar to this project as the study was also conducted on outpatient clinics for a public hospital in Singapore.

While most references shared the general consensus that no-show patient appointments are defined as patients who neither kept nor cancelled scheduled appointments, Huang and Hanauer (2014) brought up an interesting point that a cancelled appointment should be considered as no-show if it was cancelled with minimal lead time that the appointment slot cannot be filled. These findings are useful as a starting base to give us an idea of what is essential for the analysis as well as adding on to what other research studies had done. For example, the given dataset was lacking of some variables such as appointment age as seen in some of the secondary data. We can explore the data to determine if we could derive it instead. At the same time, only Huang and Hanauer (2014) accounted for the distance between the outpatient clinics and the patients' residence as being a potential factor for no-show appointment. We can compute this variable and include it for our own analysis.

3.0 METHODOLOGY



Figure 2: Flow Diagram of Modeling Process

The above figure illustrates the modeling process used for our analysis. After studying past literature papers, we proceeded to clean the data and prepared the data according to the analytical sandboxes. We have also conducted exploratory data analysis to allow us to understand more about the factors that may relate to no-show appointments. During this process, we went back to the data cleaning and preparation stage several times as we gained more insights on the variables as well as the more appropriate way to prepare some variables for the models. Once the models ran, we evaluated and assessed the performance of the models with several statistics such as Whole Model Test, Fit Statistics, Receiver Operating Characteristic (ROC) curve.

3.1 DATA PROVIDED

Our project sponsor has provided us with the data of the Children and Adolescents department’s 77,205 outpatient records with a year of data (2015 - 2016). These records are processed by the hospital staff working on the front desk. Although our project sponsor has preliminarily cleaned the data, additional data preparation work is required to allow us to focus on the key objectives and provide meaningful analysis. For example, the data includes irrelevant records of patients who were involved in a research study with medical researchers, conducted through phone.

3.2 DATA CLEANING AND DATA PREPARATION

3.2.1 Missing Data

Columns	N	N Missing
REF_TYPE	72158	3
SEX	72160	0
Revised Nationality	69956	2205
DOB	69956	2204
RACE	69956	2205
AGE	72160	0
TRT_OU_CD	72160	0
TRT_CAT	55741	16423
VISIT_NO	69956	2205
VISIT_TYPE	72160	0
VISIT_DATE	71794	366
VISIT_TIME	72160	0
PAT_CLASS	72160	0
PLAN_IND	72160	0
GROSS_AMOUNT_OTHER	69956	2204
GROSS_TAX_OTHER	69956	2204
PAYABLE_AMOUNT_OTHER	69956	2204
TAX_AMOUNT_OTHER	69956	2204
SUBSIDY_OTHER	69956	2204
ATTN_PHY	72160	0

Figure 3: Missing data

Among the variables with missing data, we have identified Race, Visit Number and Nationality as variables that are possible to fill the missing data. We are unable to fill in the missing data for Visit Date.

	Count	Number of columns missing	Patterns	RACE	VISIT_NO	Revised Nationality
1	69999	0	000	0	0	0
2	2205	3	111	1	1	1

Figure 4: Missing data pattern

Using the missing data pattern in JMP Pro, Race, Visit Number and Nationality shared the same number of missing values (2205 records). Given that each row of data is imputed whenever a patient has an appointment with Hospital X, it is possible to cross reference all records of a patient (using X_9) to find a record that is completely filled in and then fill in the rest of the missing values for the same patient.

With reference to an example, V9_1000, the missing values under Race and Nationality can be filled. However, it is not possible to fill in for the Visit Number as the column is ordinal and accumulative for each visit to Hospital X. Visit Date is cross checked with Plan IND. As it turns out, patient V9_1000 had cancelled the appointment. Thus, the missing value for Visit Number is valid as the patient did not attend the appointment. This goes the same for 2204 rows that had the missing data.

3.2.2 Duplications

For ATTN PHY, we used the recode function to rectify any duplication of names as the names could be in both lower and upper case.

3.2.3 Rectifying Discrepancies

There are some inconsistency in the gender and nationality of the patient records. Some patients are recorded as male and female for each appointment visits. We have rectified such discrepancies.

3.2.4 Extraction of dates

We have extracted the day and month from the appointment dates in order to analyse when the appointments are booked the most and has the highest no-show rate.

3.2.5 Variables Binned

With approval from the project sponsor, any patient whose age is above 25 years are excluded from the analysis as the project focuses on studying no-show appointments for child and adolescents and the number is shown to be insignificant (43 records). Age and Visit Time are binned accordingly.

3.2.6 Variables and Dimension reduction

There are 25 Reference Types from which a patient can be referred to Hospital X for an appointment. We consolidate some of the categories together as the numbers are insignificant such as combining private practitioner and private hospital into private institution. There are only 10 Reference Types as seen in Figure 4.

REF_TYPE		
Count	Old Values (21)	New Values (11)
3793	Comm MH Service	Comm MH Service
247	CommHlth Assist	
4566	Intra-Hosp A&E	Intra-Hosp
4520	Intra-Hosp Ward	
2268	Intra-Hosp SOC	
815	NHG Hosp/Inst	NHG Hosp/Inst
13060	NHG Polyclinics	NHG Polyclinics
2621	Others	Others
141	Alexandra Healt	
102	SAF	
71	Lawyer	
29	Direct SOC	
10	Natl Uni Health	
7	Step-Down Care	
6	Jurong Health	
2906	Police/MHA	Police/MHA
708	Private Institution	Private Institution
3508	School	School
3469	Self	Self
8478	Singhealth Hos	Singhealth Hos

Figure 5: Consolidating Reference Types

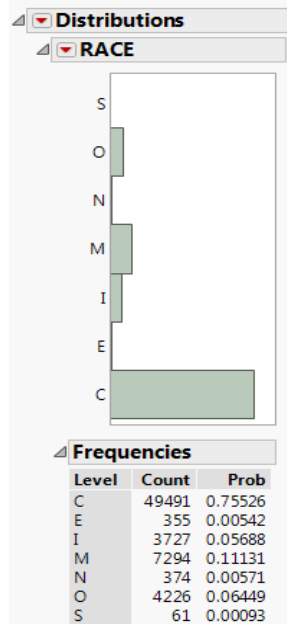


Figure 6: Distribution of Race

There are 7 race categories; M (Malay), C (Chinese), I (Indian), E (Eurasian), S (Sikh), O (Others), and N (None). We decided to combine N and O together as others could also include no race as well.

For Clinic ID, only patients who have visited either Clinic A or Clinic B are retained as the other locations are irrelevant to the project and the number is insignificant (16 records).

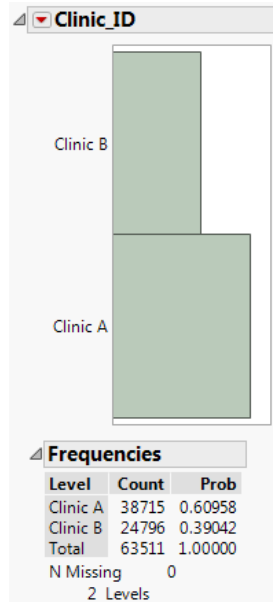


Figure 7: Distribution of Clinic ID

There are 7 categories for Visit Type. FV and RV refer to patients having an appointment with a doctor for first visit or reviewed visit respectively. AF and AR refer to patients having an appointment with an allied health professional for first visit or reviewed visit respectively. RW refers to patients visiting the clinic without arranging any appointment while XP refers to non-patients. TT includes patients involved in research study with researchers through the phone and also patients that visit the clinic to collect their prescriptions.

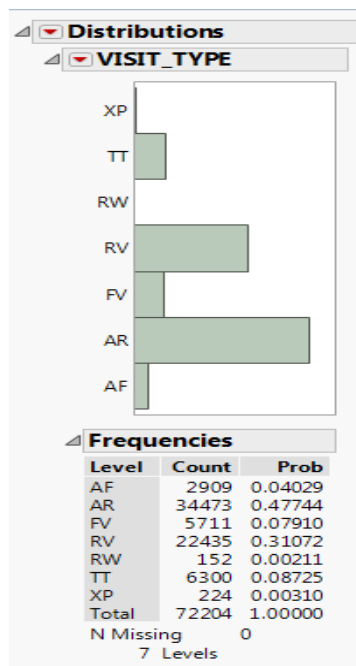


Figure 8: Distribution of Visit Type

As the focus of this project is on understanding patients' no show appointments, RW, XP and TT will be excluded from further analysis.

For ATTN PHY, there are 158 names that are attached to each patient record. Some of the names include generic names such as 'CGC Trainee MO1' and 'Medical Officer 3'. We filtered these generic names into a separate table to study if patients attached to a generic name have a higher no show rate than a real person name.

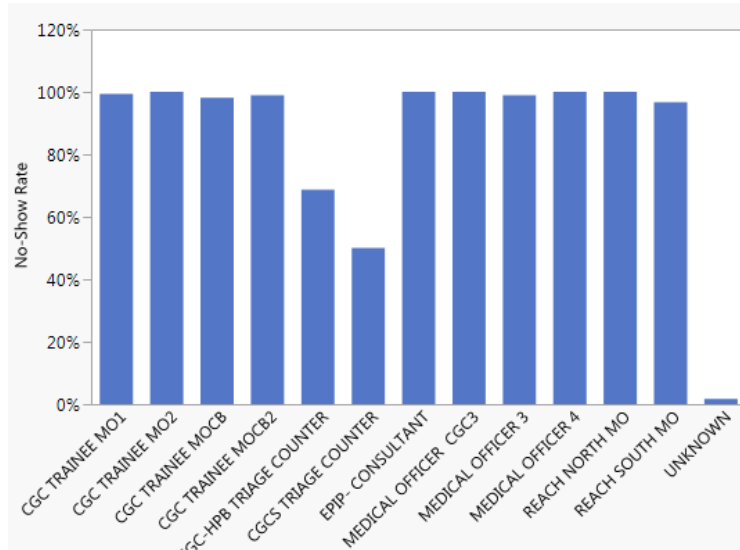


Chart 1: No-Show Rate for Generic Names

Based on the above table, there are high no-show rates for ATTN PHY with generic names. However, we have clarified with the project sponsor that the patients do not know the ATTN PHY names beforehand and thus the generic names do not affect patients' no show rate. These generic names are assigned by the staff whenever a patient does not show up or if the visit type is of a research study.

Therefore, we decided to exclude staff that attended to less than 100 patients in total and also any categories that do not specify any staff e.g. CGC Trainee MOCB or Girls' Home.

3.2.7 New Variables Derived

As mentioned earlier, the given data does not have some variables, such as appointment age, that were highlighted by other research studies. Using Visit Date, we are able to compute the appointment lead time between a patient's previous scheduled appointment and the next scheduled appointment. In addition, Clinic Switch is derived to study if there is any impact on the no-show rate of patients whose appointments are switched between the two clinics. There are 12,425 records of patients who have attended both clinics at least once.

After the data preparation process, we retained about 82% of the original data with 63,511 records left.

3.3 GEOSPATIAL DATA PREPARATION

With two different clinics situated at different parts of Singapore, we realized that there are potential insights that could be gained by heading towards the geospatial direction. It adds an additional factor that considers the distance of a patient's residence from the location of the clinic into the analysis. Maps also make it easier for us to recognize patterns that were previously buried in rows and columns.

As the data only contains the postal districts and postal codes of the patients, we need to derive the longitude and latitude points of each postal code. Other issues that arise were some patients have multiple postal codes as they have changed their residence over time and there were 2,503 records showing invalid postal district (denoted by 99).

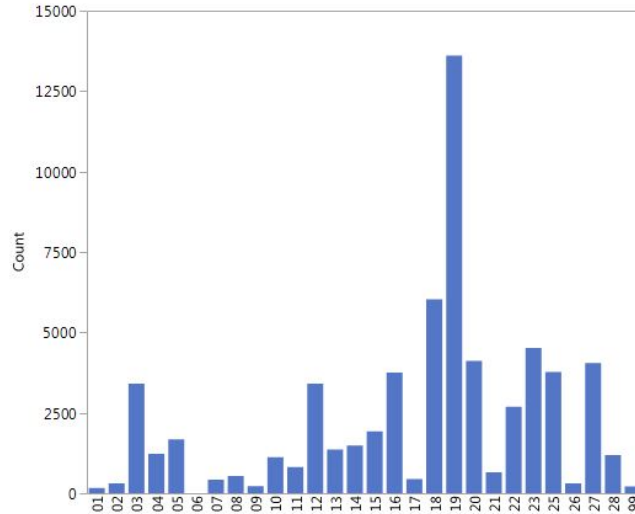


Chart 2: Distribution for Postal District

We cross-referenced patient records and managed to reduce the number of records showing invalid postal district to 216 records. We also updated the records to ensure that each patient will only have one postal code and postal district. With the advice of our project supervisor, we used Tableau 10 to generate the longitude and latitude points.

With the longitude and latitude points, we can also derive the distances of patient's residence to each clinic. Firstly, we need to convert the coordinates from World Geodetic System, WGS 84 to Singapore Coordinate System, SVY21 before using the below formula to compute the distances.

$$\left(\begin{array}{l} \text{If} \\ \text{Clinic_ID} == \text{"Clinic A"} \Rightarrow \text{Distance from Clinic (M)} \\ \text{else} \end{array} \right. \Rightarrow \text{Distance from Clinic (M)} = \sqrt[n]{ \left(33762.62 - \text{Lat (M)} \right)^2 + \left(40462.23 - \text{Long (M)} \right)^2 } \\
 \left. \begin{array}{l} \\ \\ \end{array} \right) \Rightarrow \text{Distance from Clinic (M)} = \sqrt[n]{ \left(28075.84 - \text{Lat (M)} \right)^2 + \left(28962.38 - \text{Long (M)} \right)^2 }$$

Figure 9: Formulation of Distance from Clinic

With this data preparation, we have a new variable, distance from clinic to be inputted into our models.

3.4 ANALYTICAL SANDBOXES

For modelling, we can either analyze the data of individual records as an isolated episode or analyze the data combined across the patients (grouped data by patient analysis). According to Cohen, Sanborn and Shiffrin (2008), grouping can distort the form of data, and different individuals might perform the task using different processes and parameters. However, they have shown that there are occasions where grouped analysis outperforms individual analysis. To test this literature review, we will use two sandboxes; one for analyzing each individual records and another for analyzing records grouped by patients.

The sandbox for analyzing each individual record can then be segregated further into appointments to see a doctor and appointments to see an allied health professional. The reason being is that the response variable for allied health professionals has an additional category 'cancelled appointments'. Thus, we have prepared the following for our subsequent models:

1. Per episode for doctors (0-Attended, 1-No-Show): Logistic regression and decision tree
2. Per episode for allied health professionals (0-Attended, 1-Cancelled, 2-No-Show): Multinomial regression and decision tree
3. Per patients (0-Attended, 1-Cancelled, 2-No-Show): Multiple linear regression

3.5 LOGISTIC REGRESSION MODEL

As the dependent variable, Plan IND is nominal (contains multiple categorical classes), logistic regression is selected as an appropriate modeling technique to be used. Logistic regression deals with categorical response variable by using a logarithmic transformation on the response variable which allows us to model a nonlinear association in a linear way. It is important to note that logistic regressions work with odds rather than proportion. The odds are simply the ratio of the proportions for the two possible outcomes. If y is the proportion for one outcome, then $1 - y$ is the proportion for the second outcome.

3.5.1 Dealing with Multicollinearity

As most of the data are categorical variables, we ran chi-square tests to evaluate the relationship of each variable and the dependent variable, Plan IND. This is important as logistic regression is sensitive to extremely high correlation among independent variables, which would give rise to a large standard error parameter estimates. A p -value ≤ 0.05 (as seen in Figure 10) shows that the independent variable is statistically different from the dependent variable. The chi-square tests have shown that there is at least a statistically significant relationship between each variable and the dependent variable (*Refer to Appendix 1.0 for chi-square tests of the various independent variables*).

The screenshot shows the SPSS output for a contingency analysis. It includes a table for tests and a table for measures of association.

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	234.382	<.0001*
Pearson	194.109	<.0001*

Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.0015	0.0064	-0.0112	0.0141
Kendall's Tau-b	0.0008	0.0034	-0.0060	0.0076
Stuart's Tau-c	0.0006	0.0028	-0.0048	0.0061
Somers' D C R	0.0005	0.0022	-0.0038	0.0049
Somers' D R C	0.0012	0.0054	-0.0093	0.0118
Lambda Asymmetric C R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R C	0.0000	0.0000	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000	0.0000	0.0000
Uncertainty Coef C R	0.0030	0.0003	0.0024	0.0037
Uncertainty Coef R C	0.0009	0.0001	0.0007	0.0012
Uncertainty Coef Symmetric	0.0014	0.0002	0.0011	0.0018

Figure 10: Example of a Chi-Square Test for Plan IND by Appointment Age

3.5.2 Dealing with Complete or Complete-quasi separation

When running logistic regression, we may run into a problem of a complete separation or quasi-complete separation. It occurs when a predictor variable is able to predict the response variable perfectly. E.g. Observations with $Y=0$ when all values of $A1 \leq 2$ and observations with $Y=1$ when $A2$ have values >2 . In such cases, Y separates A completely and there is no need for estimating a model as the maximum likelihood estimate for $A1$ or $A2$ does not exist. Thus, we need to make sure that the outcome variable is not a dichotomous version of a variable in the model.

3.6 DECISION TREE MODEL (RECURSIVE PARTITIONING)

Decision tree modelling is a multiple variable analysis that predicts future observations based on a set of decision rules that recursively splits independent variables into homogeneous zones. It provides unique capabilities to supplement and complement the logistic regression. Unlike logistic regression, decision tree is able to handle incomplete data and does not require any statistical assumptions concerning the data. This is prevalent in the project as some of the patients' postal codes are missing or invalid to compute any distance from the clinic.

4.0 ANALYSIS

4.0.1 Age Group

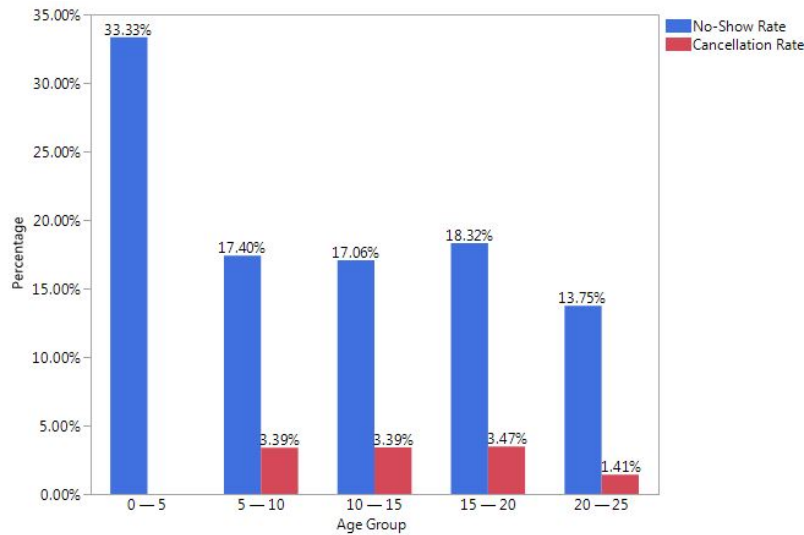


Chart 3: No-Show Rate for Age Group

No-show rate is calculated as a percentage of number of no-shows over the total number of patients that fall in a particular category. According to Michael et al (2016), no-show rate generally decreases as age group increases. Based on the above chart, the overall average no-show rate decreases with age group until the age of 15-20 when there is an increase before dropping down again. The highest no-show rate is the age group of 0-5.

4.0.2 Visit Type

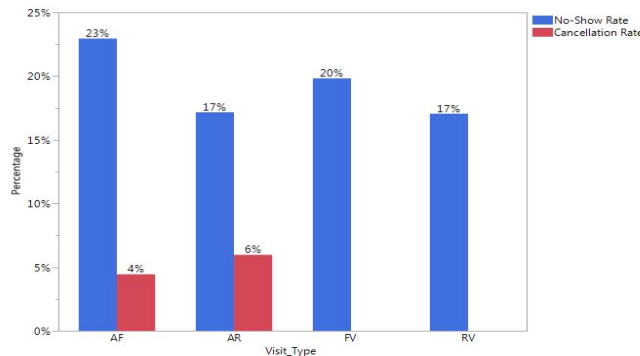


Chart 4: No-Show Rate for Visit Type

From the Visit Type chart, it can be observed that the first visits to doctors and allied health professionals have higher no-show rates than reviewed visits. There is no cancellation rate for FV and RV as the staff, operating under the doctors, combined the cancelled appointments and no-show appointments as no-show appointments.

4.0.3 AF and AR by Time of Day

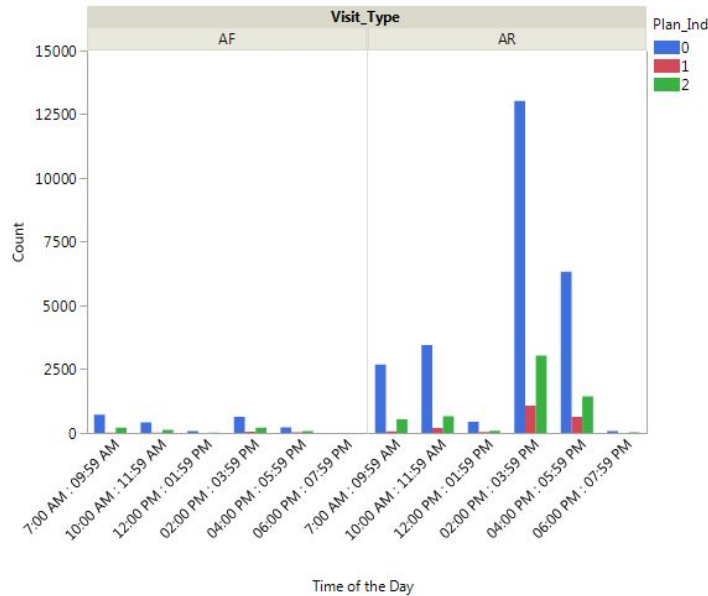


Chart 5: Distribution of PLAN_IND for AF & AR Throughout The Day

There are more reviewed visits scheduled as compared to first visits to allied health professionals. Patients have a stronger preference for reviewed visits in the late afternoon as evident by the number of number of appointment scheduled and the high attendance.

4.0.4 FV and RV by Time of Day

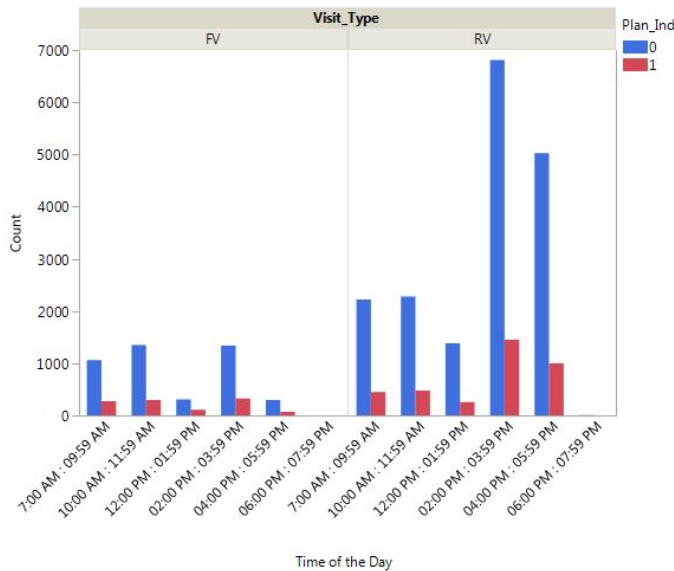


Chart 6: Distribution of PLAN_IND for FV & RV Throughout The Day

Similarly, there are more reviewed visits to doctors as compared to first visits to doctors. Patients have a stronger preference for reviewed visits in the late afternoon as evident by the number of number of appointment scheduled and the high attendance.

This analysis can be strengthened further if we are able to crosscheck the doctors and allied health professionals' work schedules.

4.0.5 Appointment Age

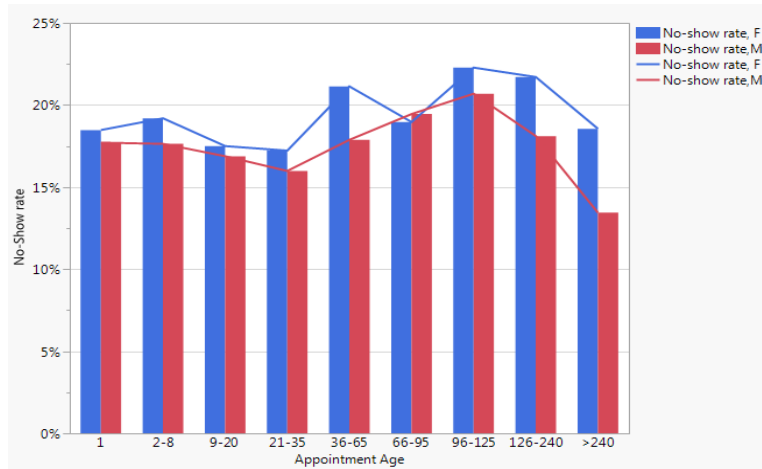


Chart 7: No-Show Rate for Appointment Age

The no-show rate for males with respect to appointment age was lower than females. According to Michael et al (2016), past research has shown that no-show rate will increase as appointment age increases. The above chart does not reflect it as the no-show rate varies as appointment age increases. One explanation for this nonconformity could be due to the one year dataset. More data is needed for this analysis to be effective.

4.0.6 Postal District

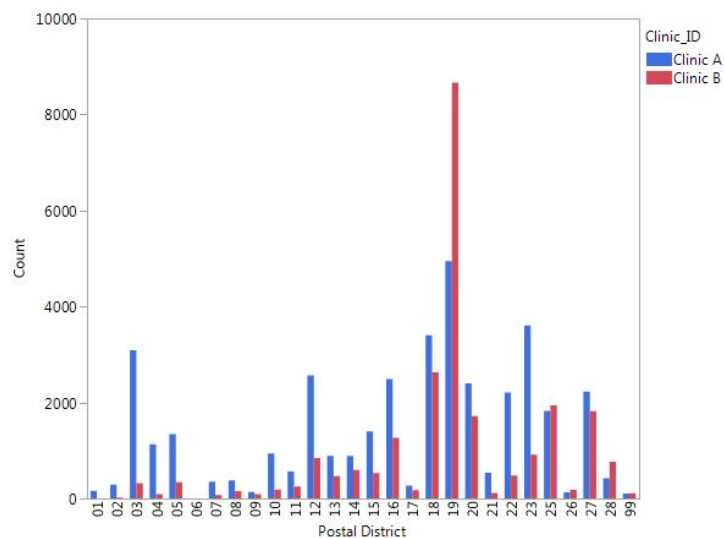


Chart 8: Postal District by Clinic Location

The postal district showed that the main bulk of the patients, in the dataset, resided in District 19. District 19 consists of general location around Serangoon Garden, Hougang and Punggol. Clinic B has a significant number of patients from District 19 due to the close proximity

of its location. The below chart depicts the distribution of patients living in each postal district around Singapore. The most densely populated district (highlighted in red) is District 19.

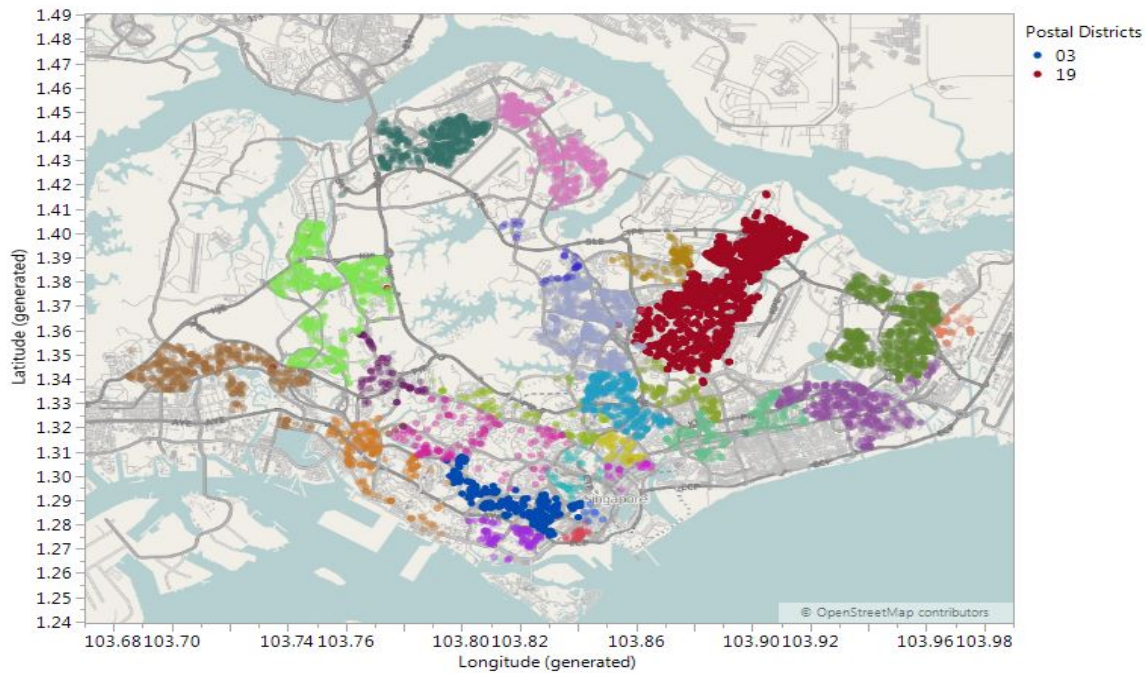


Figure 11: Distribution of Patients in Each Postal District

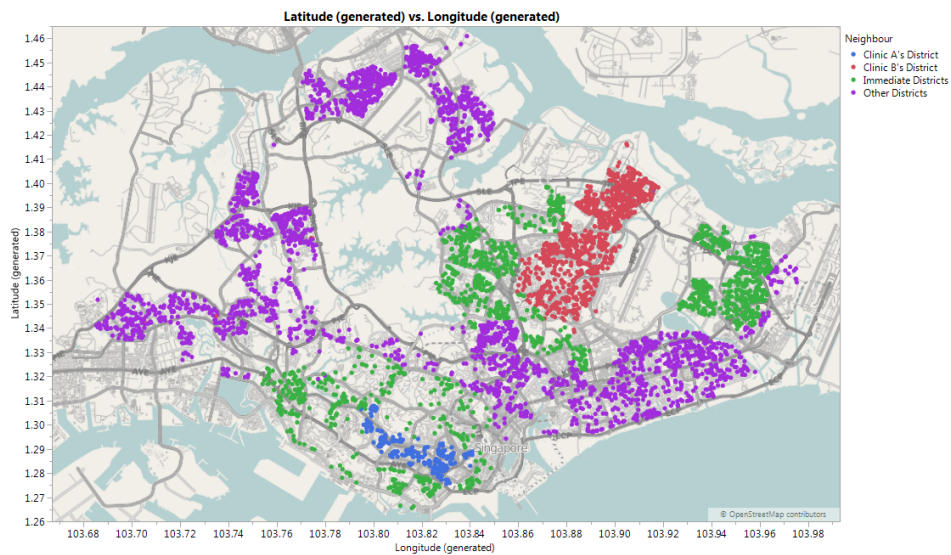


Figure 12: Distribution of Patients around Singapore

In order to understand the distribution of patients by the clinic locations, we grouped the postal districts into the districts that the clinics are located in, the next immediate districts and other districts. As seen in Figure 12, Clinic A is located in district 3 (highlighted in blue) while clinic B is located in district 19 (highlighted in red). Districts 1, 2, 4, 5, 6, 9, 10, 13, 18, 20 and 28 (highlighted in green) are the immediate neighbours around the respective clinic's district. The other districts, which are considered further apart from either of the clinic locations, are highlighted in purple. Figure 12 depicts a high density of patients living in Clinic B's district, which explained the large portion of patients from District 19 (as seen in chart 8) having appointments in Clinic B.

4.0.7 Distance from Clinic

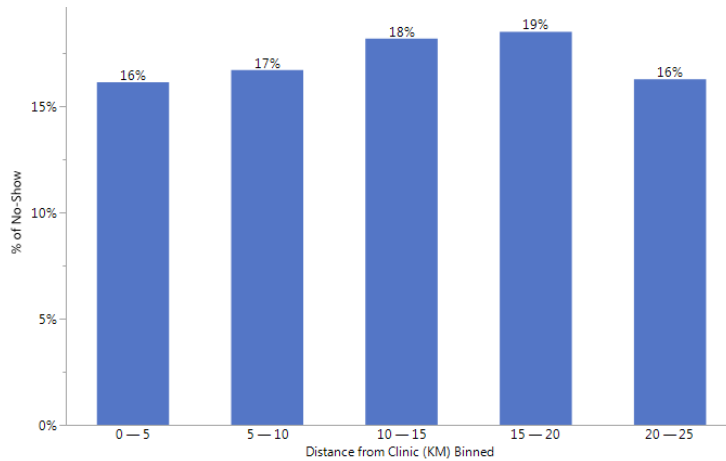


Chart 9: No-Show Rate for Distance from Clinic (KM)

The above chart shows that as the no-show rate increases with an increase in the distance from the clinic, with the exception of the distance of 20km to 25km where there is dip in no-show rate. This may be attributed to 981 transaction records whose postal code is unavailable or invalid and thus is not part of this analysis. The no-show rate of the unaccounted records is 25.68%.

4.0.8 Clinic Switch

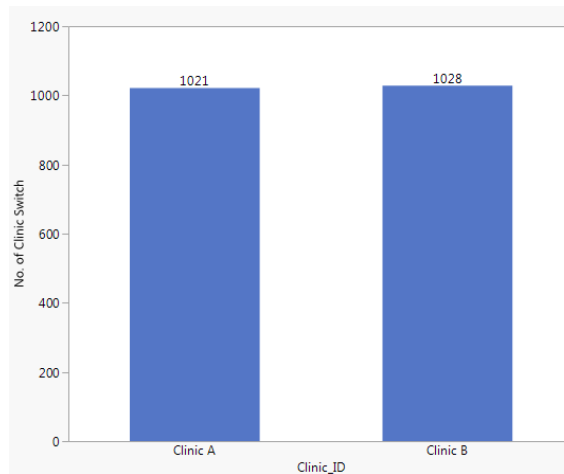


Chart 10: Number of Clinic Switch

The number of patients whose appointments are switched between the two clinics is almost evenly distributed. One explanation could be that the appointments are being scheduled according to the doctors and psychologists' clinic schedules (who may move between the two clinics) instead of the patients' clinic preference. The chart below showed that patients who were switched to clinic B have a higher no-show rate than clinic A.

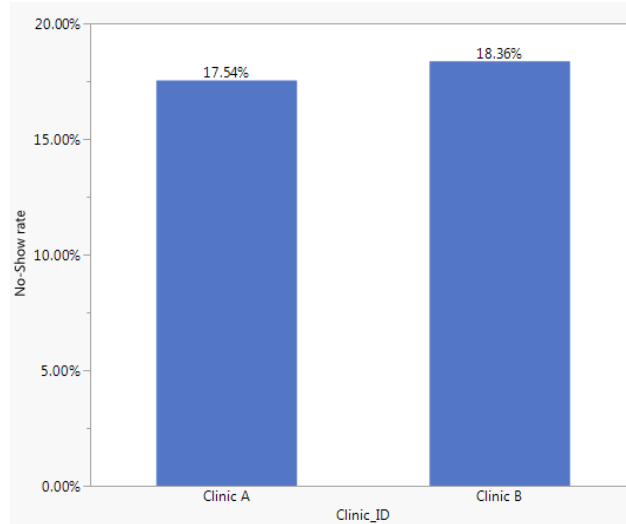


Chart 11: No-Show Rate for Number of Clinic Switch

4.1 EVALUATION OF LOGISTIC REGRESSION MODEL

4.1.1 Per episode for Doctors

Whole Model Test				
Model	-LogLikelihood	DF	ChiSquare	Prob> ChiSq
Difference	310.4773	50	620.9546	<.0001*
Full	8784.8274			
Reduced	9095.3047			
RSquare (U)		0.0341		
AICc		17671.9		
BIC		18074		
Observations (or Sum Wgts)		19714		

Figure 13: Whole Model Test for Logistic Regression Model (Doctors)

The whole model test is testing:

H_0 : The logistic model is NOT useful

H_1 : The logistic model is useful

Decision: Reject H_1 if the p value is NOT significant

Since the above results showed that p value $<.0001$, there is sufficient statistical evidence to reject the null hypothesis. The logistic model is useful to explain the odds of no-show patient appointments. In other words, the overall model is significant at the 0.001 level according to the Model chi-square statistic.

Lack Of Fit			
Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	19084	8679.8088	17359.62
Saturated	19134	105.0186	Prob> ChiSq
Fitted	50	8784.8274	1.0000

Figure 14: Lack of Fit Test for the logistic regression model (Doctors)

The lack of fit test is used to evaluate if our model is adequate in explaining the odds of no-show patient appointments. The lack of fit test is testing the hypothesis of the following:

H_0 : The logistic model is adequate

H_1 : The logistic model is inadequate, i.e. there is lack of fit

Decision: Reject H_1 if the p value is large and NOT significant

The above figure shows that lack of fit chi-square is insignificant (Prob>Chisq = 1.0000) and supports the conclusion that there is little to be gained by introducing additional variables.

Effect Likelihood Ratio Tests				
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Race	5	5	185.612313	<.0001*
Nationality	2	2	6.56937748	0.0375*
Gender	1	1	3.88746987	0.0486*
Age	1	1	3.30211196	0.0692
Clinic_ID	1	1	67.2238973	<.0001*
Visit_Type	1	1	4.14189034	0.0418*
Pat_Class	2	2	36.3774238	<.0001*
Month	11	11	60.7451503	<.0001*
Day	4	4	13.595243	0.0087*
Neighbour	3	3	5.36073393	0.1472
Distance from Clinic (KM) Binned	4	4	2.90220149	0.5743
Ref_Type2	6	6	38.6745663	<.0001*
Appointment Age (Binned)	9	9	68.0088784	<.0001*

Figure 15: LRT for Logistic Regression Model (Doctors)

While SAS JMP Pro provides the parameter estimates based on the Wald's test, we used the likelihood-ratio tests (LRT) to assess the individual parameters. LRT are more reliable than Wald's test as it is computed iteratively. In this case, not all predictors are significant. Age, Neighbour, Distance from Clinic are insignificant while Race, Clinic Id, Patient Class, Month, Reference Type and Appointment Age have p -value<0.0001. The parameter estimates report gives more detailed information on which category is significant within each predictor variable (Refer to Appendix 2.0 for parameter estimates report).

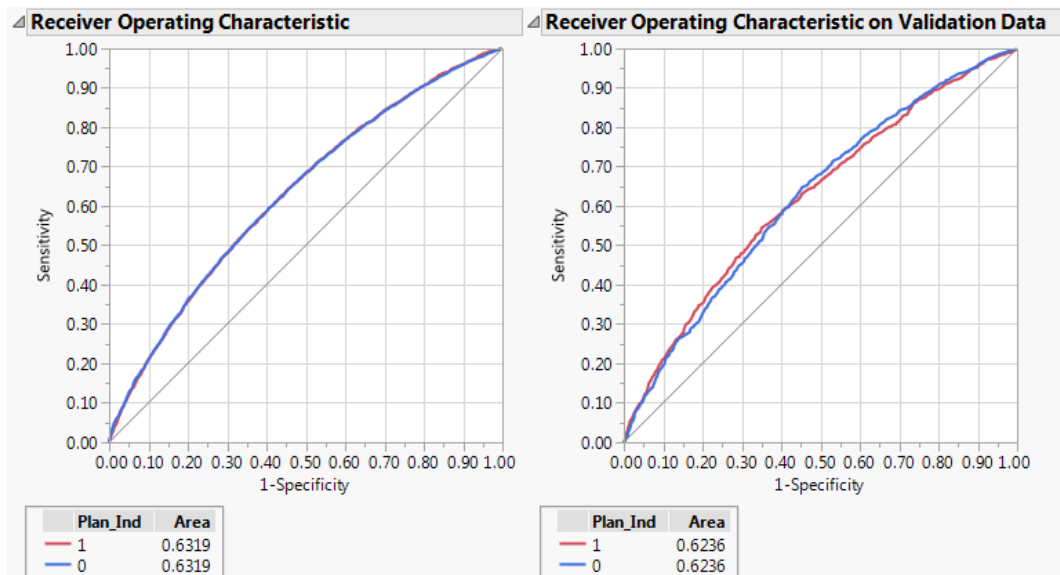


Figure 16: ROC Curve for Logistic Regression Model (Doctors)

The ROC curve, which is a plot of sensitivity by (1-specificity) for each value of x, indicates low distinguish ability (not a very good model yet the model can be used).

Confusion Matrix					
Training			Validation		
Actual	Predicted Count		Actual	Predicted Count	
Plan_Ind	1	0	Plan_Ind	1	0
1	14	3406	1	6	1153
0	10	16284	0	10	5461

Figure 17: Confusion Matrix for Logistic Regression Model (Doctors)

<u>True Negatives</u> (Actual 0, Predict 0)	<u>False Positives</u> (Actual 0, Predict 1)
16,285	10
<u>False Negatives</u> (Actual 1, Predict 0)	<u>True Positives</u> (Actual 1 Predict 1)
3,406	14

Table 1: Contingency Table for Logistic Regression Model (Doctors)

In order to assess the overall performance of the model, we look into the misclassification rate, which can be calculated by the following:

$$\text{Misclassification rate} = (\text{False positives} + \text{False negatives}) \div \text{Total} = 17.33\%$$

Therefore, the model predicts 82.67% of the patient appointments attendance correctly. However, the model is only able to predict 0.41% of the no-show appointments.

Since the data has a large proportion of appointments that was attended as compared to no-show appointments, it would not be right for the model to calculate the cut-off rate to be 0.5. Thus, we need to impute a new cut-off rate to gauge the probability of no-show appointments better.

$$\text{If } \begin{cases} \text{Prob}[1] \geq 0.15 & \Rightarrow \text{[New] Most Likey Plan_IND} = 1 \\ \text{else} & \Rightarrow \text{[New] Most Likey Plan_IND} = 0 \end{cases}$$

Figure 18: Formulation for Cut-off rate (15%)

Contingency Table				
		Plan_Ind		
Count		0	1	Total
Total %	Col %			
	Row %			
0		7709	995	8704
		39.10	5.05	44.15
		47.31	29.09	
		88.57	11.43	
1		8585	2425	11010
		43.55	12.30	55.85
		52.69	70.91	
		77.97	22.03	
Total		16294	3420	19714
		82.65	17.35	

Figure 19: Contingency Table (15%) for Logistic Regression Model (Doctors)

<u>True Negatives (Actual 0, Predict 0)</u>	<u>False Positives (Actual 0, Predict 1)</u>
7,709	8,585
<u>False Negatives (Actual 1, Predict 0)</u>	<u>True Positives (Actual 1 Predict 1)</u>
995	2,425

Table 2: Contingency Table (15%) for Logistic Regression Model (Doctors)

The misclassification rate for 15% cut-off rate is 48.59%. Therefore, the model predicts 50.72% of the patient appointments attendance correctly. Based on true positives, the model is now able to predict 70.12% of the no-show patient appointments attendance correctly. In order to decide on the optimal cut-off rate, we repeated the computation of true positives for cut-off rate of 10%, 16%, 17%, 18%, 19%, 20%.

Summary of the various cut-off rate		
Cutoff (%)	No-show Prediction (%)	Model Prediction (%)
10	95.56	25.99
15	70.12	50.72
16	64.44	56.06
17	58.13	60.13
18	52.54	63.72
19	47.63	63.72
20	42.40	69.00

Table 3: Summary of The Various Cut-off Rate for Logistic Regression Model (Doctors)

4.1.2 Per episode for Allied Health Professionals

A similar logistic regression model for allied health professionals was conducted with the Plan IND having three categories (Attended, Cancelled, No-show). A summary of the various cut-off rate is shown below. (Refer to Appendix 3.0 for the complete evaluation of logistic regression model for allied health professionals)

Summary of the various cut-off rate		
Cutoff (%)	No-show Prediction (%)	Model Prediction (%)
10	99.21	18.84
15	72.10	45.14
16	64.18	51.14
17	56.78	56.16
18	50.79	60.00
19	44.52	63.32

20	39.27	65.80
----	-------	-------

Table 4: Summary of the Various Cut-off Rates for Logistic Regression Model (Allied Health Professionals)

4.2 EVALUATION OF DECISION TREE MODEL

4.2.1 Per episode for Doctors

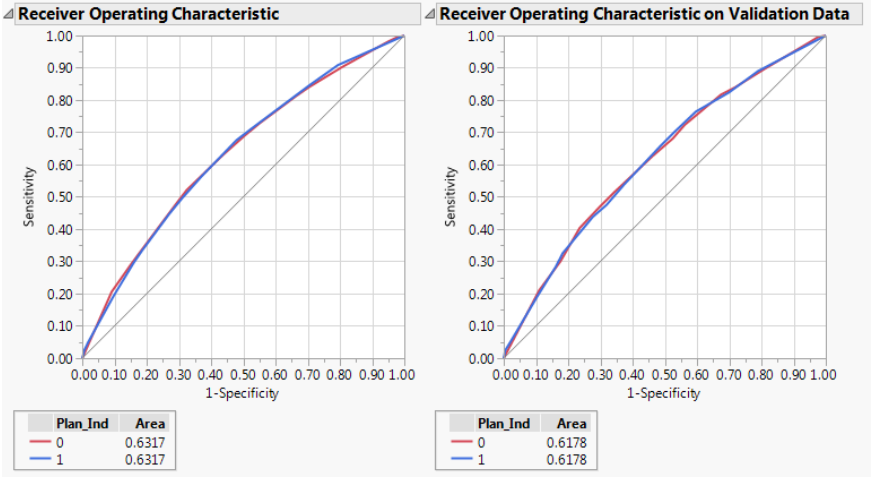


Figure 20: ROC Curve for Decision Tree Model (Doctors)

The ROC curve for the decision tree model indicates a low distinguish ability (not a very good model, yet the model can be used) in identifying no-show appointments from the model. The decision tree model has an almost similar distinguishing power to that of the logistic regression for doctors.

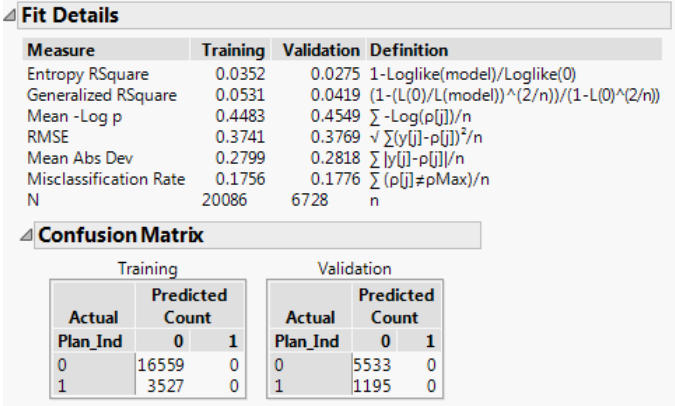


Figure 21: Fit Details & Confusion Matrix for Decision Tree Model (Doctors)

The misclassification rate for the decision tree is 17.56. Therefore, the model predicts 82.44% of the patient appointments attendance correctly. Based on the true positives, the decision tree model predicts 0% of the no-show appointments. Similarly to the logistic regression, it would not be right for the model to calculate the cut-off rate at 0.5. Thus, we need to impute a new cut-off rate to gauge the probability of no-show appointments better.

Contingency Table			
Plan_Ind			
Count	0	1	Total
Total %			
Col %			
Row %			
0	8634 42.99 52.14 88.31	1143 5.69 32.41 11.69	9777 48.68
1	7925 39.46 47.86 76.87	2384 11.87 67.59 23.13	10309 51.32
Total	16559 82.44	3527 17.56	20086

Figure 22: Contingency Table (15%) for Decision Tree Model (Doctors)

<u>True Negatives (Actual 0, Predict 0)</u>	<u>False Positives (Actual 0, Predict 1)</u>
8,634	7,925
<u>False Negatives (Actual 1, Predict 0)</u>	<u>True Positives (Actual 1 Predict 1)</u>
1,143	2,384

Table 5: Contingency Table (15%) for Decision Tree Model (Doctors)

The misclassification rate for 15% cut-off rate is 45.15%. Therefore, the model predicts 54.85% of the patient appointments attendance correctly. Based on true positives, the model is now able to predict 67.59% of the no-show patient appointments attendance correctly. In order to decide on the optimal cut-off rate, we repeated the computation of true positives for cut-off rate of 10%, 16%, 17%, 18%, 19%, 20%.

Summary of the various cut-off rate		
Cutoff (%)	No-show Prediction (%)	Model Prediction (%)
10	90.87	32.97
15	67.59	54.85
16	67.59	54.85
17	67.59	54.85
18	67.59	54.85
19	56.33	61.84
20	42.40	69.00

Table 6: Summary of the Various Cut-off rates for Decision Tree Model (Doctors)

The optimal cut-off rate for the logistic regression model (doctors) is 19%. Under the decision tree model, the significant variables are Race, Clinic Id, Distance from the Clinic, Appointment Age, Patient Class, Visit Type, Month, Age and Reference Type (*Refer to Appendix 4.0 for the column contribution*).

4.2.2 Per episode for Allied Health Professionals

A similar decision tree model for allied health professionals was conducted with the Plan IND having three categories (Attended, Cancelled, No-show). A summary of the various cut-off rate is shown below. (Refer to Appendix 5.0 for the complete evaluation of decision tree model for allied health professionals)

Summary of the various cut-off rate		
Cutoff (%)	No-show Prediction (%)	Model Prediction (%)
10	100.00	17.74
15	64.74	46.79
16	42.62	67.47
17	42.62	67.47
18	42.62	67.47
19	42.62	67.47
20	42.62	67.47

Table 7: Summary of the Various Cut-off Rate for Decision Tree Model (Allied Health Professionals)

In this case, there is no optimal cut-off rate as the model would either go lower than 50% prediction for either no-show appointments or the patient appointment attendance. Under the decision tree model, the significant variables are Race, Visit Time, Age, Reference Type, Visit Type, Neighbour, Gender, Month and Appointment Age (Refer to Appendix 6.0 for the column contribution).

4.3 COMPARISON OF LOGISTIC REGRESSION AND DECISION TREE MODELS

4.3.1 Model comparison for doctors

Model Comparison												
Predictors												
Measures of Fit for Plan_Ind												
Validation	Creator	.2	.4	.6	.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Training	Fit Nominal Logistic					0.0341	0.0515	0.4456	0.3725	0.2774	0.1733	19714
Training	Partition					0.0352	0.0531	0.4483	0.3741	0.2799	0.1756	20086
Validation	Fit Nominal Logistic					0.0302	0.0457	0.4494	0.3742	0.2790	0.1754	6630
Validation	Partition					0.0275	0.0419	0.4549	0.3769	0.2818	0.1776	6728

Figure 23: Model Comparison for Doctors

In terms of misclassification rate, the logistic regression model has a misclassification rate of 17.33%, whereas the decision tree model has a misclassification rate of 17.56%. As the misclassification rate for logistic regression model is slightly lower than decision tree model, it can be concluded that the logistic regression model is able to predict the patient appointments slightly more correctly.

Predictive Performance Metrics for Doctors' Models		
Metric	Logistic Regression	Decision Tree
Misclassification Rate	17.33%	17.56%
Specificity Rate	99.94%	100%
Sensitivity Rate	0.41%	0%
ROC	0.6319	0.6317

Table 8: Performance Metrics for Doctors' Models (Based on Default JMP Pro Prediction Formula)

In terms of the confusion matrix, logistic regression model has a slightly lower specificity rate of 99.94% as compared to the decision tree model at 100%. Specificity (True Negative / True Negative + False Positive) is the true negative rate. It answers the question, "If the model predicts a negative event, what is the probability that it really is negative?" In this case, the negative event refers to the patient attending his scheduled appointment. This shows that the decision tree model has a higher negative rate and thus, is able to predict the patient appointment attendances correctly.

On the other hand, logistic regression model has a higher sensitivity rate of 0.41% as compared to decision tree model at 0%. Sensitivity (True Positive / True Positive + False Negative) is the true positive rate. It answers the question, "If the model predicts a positive event, what is the probability that it really is positive?" In this case, the positive event refers to the no-show appointment. In other words, logistic regression model is better at predicting no-show appointments correctly than decision tree model.

Comparing the ROC curve, the logistic regression model has an area of 0.6319 for no-show appointments, whereas the decision tree model has an area of 0.6317 under the curve. While both models have a low distinguishing ability in identifying no-show appointments, the logistic regression model performs slightly better.

The significant factors, which were identified by both models, for no-show appointments were Race, Clinic Id, Patient Class, Month, Reference Type and Appointment Age.

Therefore, we can conclude the logistic regression model would be a better overall fit for analysing no-show appointments under doctors if we based the models on the default JMP Pro prediction calculations.

Predictive Performance Metrics		
Metric	Logistic Regression	Decision Tree
Optimal Cut-off Rate	17.00%	19.00%
Misclassification Rate	39.87%	38.16%
Specificity Rate	60.50%	63.02%
Sensitivity Rate	58.13%	56.34%

Table 9: Performance Metrics for Doctors' Models (Based on Optimal Cut-off Rate)

In terms of optimal cut-off rate, decision tree model fared slightly better than logistic regression model in terms of predicting the overall appointments' attendance (lower misclassification rate of 38.16%) and predicting attended appointments (higher specificity rate at 63.02%) correctly. On the other hand, logistic regression fared better than decision tree model in predicting no-show appointments (higher sensitivity rate of 58.13%) correctly.

4.3.2 Model comparison for Allied Health Professionals

Model Comparison												
Predictors												
Measures of Fit for Plan_Ind												
Validation	Creator	.2	.4	.6	.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Training	Fit Nominal Logistic					0.0299	0.0526	0.6351	0.4466	0.3600	0.2276	26971
Training	Partition					0.0627	0.1098	0.6345	0.4459	0.3589	0.2254	27629
Validation	Fit Nominal Logistic					0.0224	0.0395	0.6392	0.4466	0.3597	0.2247	8858
Validation	Partition					0.0546	0.0959	0.6391	0.4461	0.3590	0.2233	9068

Figure 24: Model Comparison for Allied Health Professionals

In terms of misclassification rate, the logistic regression model has a higher misclassification rate of 22.76% as compared to the decision tree model at 22.54%. Thus, the decision tree model can slightly better predict patient appointments correctly.

Predictive Performance Metrics		
Metric	Logistic Regression	Decision Tree
Misclassification Rate	22.76%	22.54%
Specificity Rate	99.95%	81.17%
Sensitivity Rate	0.18%	0%
ROC	0.6100	0.5912

Table 10: Performance Metrics for Allied Health Professionals' Models (Based on Default JMP Pro Prediction Formula)

In terms of confusion matrix, the decision tree model has a lower specificity rate of 81.17% as compared to the logistic regression model at 99.95%. Thus, the logistic regression can better predict patient appointments under allied health professionals correctly. Logistic regression model also has a higher sensitivity rate of 0.18% as compared to the decision tree model at 0%. Thus, the logistic regression has a higher true positive rate and can better predict no-show patient appointments correctly.

Comparing the ROC curve, the logistic regression (0.6100) fared slightly better than the decision tree model (0.5912) in terms distinguishing ability in identifying no-show appointments.

The significant factors, which were identified by both models, for no-show appointments were Race, Reference Type, Visit Type, Visit Time, Month, Neighbour and Appointment Age.

Therefore, we can conclude the logistic regression model would be a better overall fit for analysing no-show appointments under allied health professionals due to its higher sensitivity rate based on default JMP Pro prediction calculations.

Predictive Performance Metrics		
Metric	Logistic Regression	Decision Tree
Optimal Cut-off Rate	17.00%	16.00%
Misclassification Rate	43.84%	37.53%
Specificity Rate	59.56%	71.81%
Sensitivity Rate	44.62%	32.13%

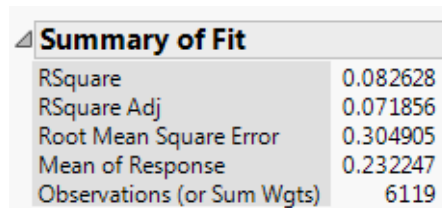
Table 11: Performance Metrics for Allied Health Professionals' Models (Based on Optimal Cut-off Rate)

In terms of optimal cut-off rate, decision tree model fared slightly better than logistic regression model in terms of predicting the overall appointments' attendance (lower misclassification rate of 37.53%) and predicting attended appointments (higher specificity rate at 71.81%) correctly. On the other hand, logistic regression fared better than decision tree model in predicting no-show appointments (higher sensitivity rate of 44.62%) correctly.

4.4 LEAST PARTIAL SQUARE REGRESSION MODEL (GROUPED DATA)

In this analysis, we will study the data by grouping the records according to the 8,200 patients. Thus, a regression model will be developed to explain no-show appointments of patients based on the same independent variables used in the previous models. The aim is to evaluate the effect of a grouped analysis.

Dummy variables (k-1) for each categorical variable are created for the regression model. The records are then grouped according to each patient before running the model.

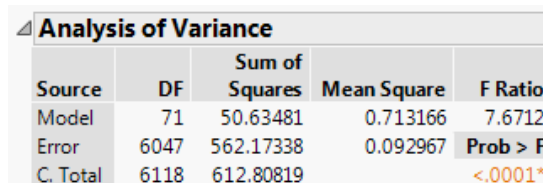


Summary of Fit	
RSquare	0.082628
RSquare Adj	0.071856
Root Mean Square Error	0.304905
Mean of Response	0.232247
Observations (or Sum Wgts)	6119

Figure 25: Fit Details for Grouped Data

The R^2 and Adjusted R^2 of the model are 0.082628 and 0.071856 respectively. The value of R^2 tells us that the independent variables can account for 8.26% of the variation in the no-show patient appointments. This confirms that there is loss of information when records are grouped by patients. As a result, the grouped analysis is subjected to noise and bias that produce distortion. Therefore, it is incorrect to assume that the relationships existing at one level of the individual analysis will necessarily demonstrate the same strength at the level of the grouped analysis. For example, a patient may have attended only 6 out of his/her 8 appointments and the appointments have been scheduled on different days and at different times. The model is unable to specify or differentiate the day and time that the no-show appointments had occurred from the other appointments.

In this case, it may be more appropriate to consider using time series and smoothing methods for the grouped data by patient analysis. These methods forecast no-show events based on past events by using stochastic models (Adel et al, 2011).



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	71	50.63481	0.713166	7.6712	
Error	6047	562.17338	0.092967		Prob > F
C. Total	6118	612.80819			<.0001*

Figure 26: Analysis of Variance for Grouped Data

The analysis of variance shows that the F-ratio is 7.6712 which is significant at $p < 0.0001$. This result tells us that there is less than 0.01% chance that an F-ratio as observed will happen if the null hypothesis is true. Therefore, the regression model result is significantly better prediction of no-show patient appointments than if the mean value of no-show patient appointments is used.

5.0 DISCUSSION

Based on the default JMP Pro Prediction formula, the results have shown that logistic regression model is a more suitable overall fit for analysing no-show appointments under doctors or allied health professionals as it has a higher sensitivity rate. It is important to note that

JMP Pro prediction formula assumes the probability of an event from categorical dependent variable occurring to be 0.5. In the context of no-show appointments, the probability of no-show appointments is much lower than 0.5. Thus, it is necessary to reformulate the prediction probability in order obtain a more realistic model prediction.

While the models have shown significant predictor factors for no-show appointments, we believe that the current iteration of models can be strengthened further in terms of its explanatory and predictive ability. The analysis may benefit by using at least 5 years' worth of data instead of just one year. The predictor variables used were the ones available to us in the given dataset and we have tried to derive other relevant variables to be included in the models. As seen in our literature review, there are other possible contributing factors to no-show appointments such as financial debt and the number of people in the household of the patient (Huang & Hanauer, 2014). This is very relevant to the project as the patients are below the age of 25 years old and thus it is very likely that an adult will be accompanying the patient. Thus, the information could be interesting to assess how the different members in a household play a part in the chances on a patient not showing up for his/her appointment.

Another limitation was the inconsistent recording of the patient records by the frontline staff. The variable, Primary Diagnostic Condition would probably have been an interesting factor to study in relation to no-show appointments. However, it has a significant portion of missing data. Hospital X should consider standardizing certain recording procedures across different departments such as recording down the cancellation of appointments and the lead time between a cancelled appointment and the scheduled appointment.

13.0 CONCLUSION

We have identified factors that are related to no-show appointments and have shown that the performance of a model may vary according to appointments by doctors or by allied health professionals. The common predictor variables among the models are Race, Reference Type, Month and Appointment Age. Hospital X can make use of our analysis as a base to improving no-show appointments. To increase the value of the models, it needs to collect more relevant predictor variables as well as increasing the data size in future analysis.

In addition, Hospital X may consider collecting data on the effectiveness of its current appointment reminder procedure. An example would be to run a simulation comparing the current appointment scheduling procedure versus new strategies. This could be done using Excel Macros (as utilised by Huang & Hanauer, 2014) to run a Monte Carlo simulation by following Muthuraman and Lawley's (2008) method (implemented by Daggy et al., 2010) that takes as input costs of patient waiting, revenue per patient, etc. as input in simulating a clinic setting. From this, a scheduling algorithm integrated with the clinic's administrative database can feasibly be designed, which can determine when and how to schedule each new appointment based on the patient's no-show probability. While these were beyond the scope of our project to conduct, they are future approaches beyond modelling which Hospital X can consider.

ACKNOWLEDGEMENT

This paper is done as part of Analytics Practicum project and we would like to thank Professor Kam Tin Seong (Associate Professor of Information Systems; Senior Advisor, SIS) for his valuable insights as well as his guidance throughout the project. We also would like to thank our project sponsor for graciously sponsoring the project, providing us with access to Hospital X's data as well as dedicating his time to ensure that we are familiar with the healthcare domain.

CONTACT INFORMATION

Your comments and questions are valued and encourage. Contact the authors at:

Name: Loh Yan Zoey

E-mail: zoey.loh.2013@socsc.smu.edu.sg

Name: Mirania Aishwarya Agarwal

E-mail: miraniaa.2013@business.smu.edu.sg

Name: Nasrullah Bin Khairullah

E-mail: nasrullahk.2013@business.smu.edu.sg

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

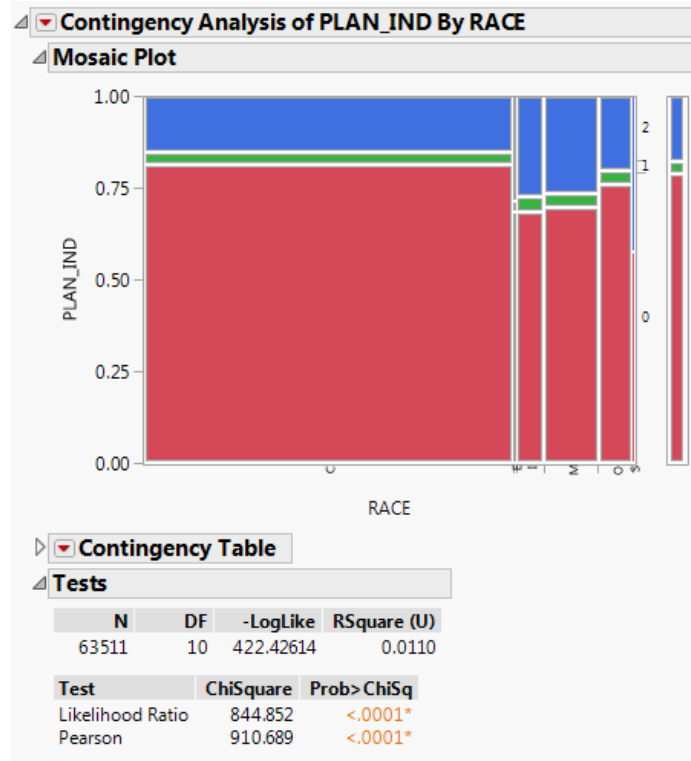
REFERENCES

- Allaeddini, A., Yang, K., Reddy, C. & Yu, S. (2011, February). *A Probabilistic Model for Predicting The Probability of No-Shows in Hospital Appointments.*
- Cohen. A. L., Sanborn. A. N., Shiffrin. R. M., (2008). *Model Evaluation Using Grouped or Individual Data.*
- Clark. W. A. V., Avery. K. L. (1975, October). *The Effects of Data Aggregation in Statistical Analysis*
- Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P. C., ... & Sands, L. (2010). *Using no-show modeling to improve clinic performance.* Health Informatics Journal, 16(4), 246-259.
- Huang, Y., & Hanauer, D.A. (2014, September). *Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach.*
- Ma. N. L., Seemanta. K., Wu. D., Ng. S. S. Y. (2014). *Predictive Analytics for Outpatient Appointments.*
- Michelle. K. (2011, January). *When Absence Speaks Louder than Words: An Object Relational Perspective on No-Show Appointments.*
- Michael. L. D., Rachel. M. G., Jerrold. H. M., Robert. J. M., Keri. L. R., Youxu. C. T., Dominic. L. V., (2016, February). *Large-Scale No-Show Patterns and Distributions for Clinic Operational Research.*
- Molfenter. T. (2013). *Reducing Appointment No-Shows: Going From Theory to Practice.*
- Muthuraman, K., & Lawley, M. (2008). *A stochastic overbooking model for outpatient clinical scheduling with no-shows.* *lie Transactions*, 40(9), 820-837.
- Naomi. L. L., Audrey P, Matthew. D. R., Bruce. L. (2004). *Why We Don't Come: Patient Perceptions on No-Shows.*
- William. S. M, M.S.W., BCD (2001). *Why They Don't Come Back: A Clinical Perspective on The No-Show Client.*
- Woo. BS., Ng. TP., Fung. DS., Chan. YH., Lee.YP., Koh. JB., Cai. Y. (2007). *Emotional and Behaviorual Problems in Singaporean Children Based on Parent, Teacher and Child Reports.*

APPENDICES

Appendix 1.0 Relationship of Variables and PLAN_IND

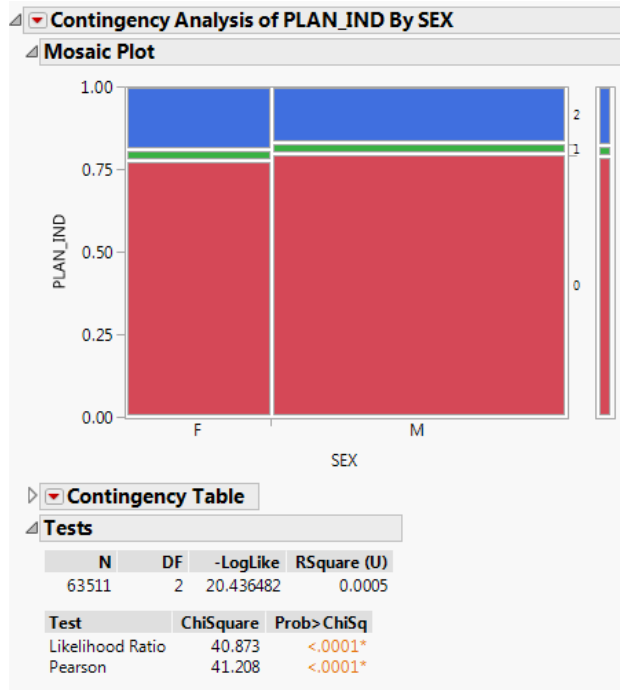
Race



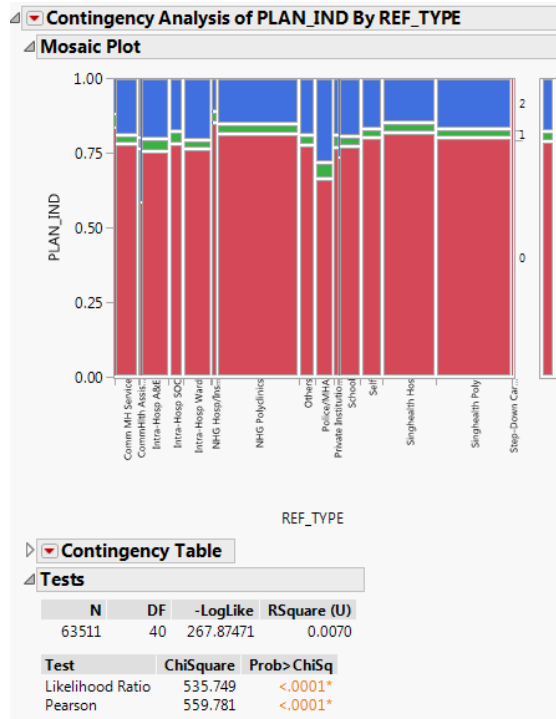
Nationality



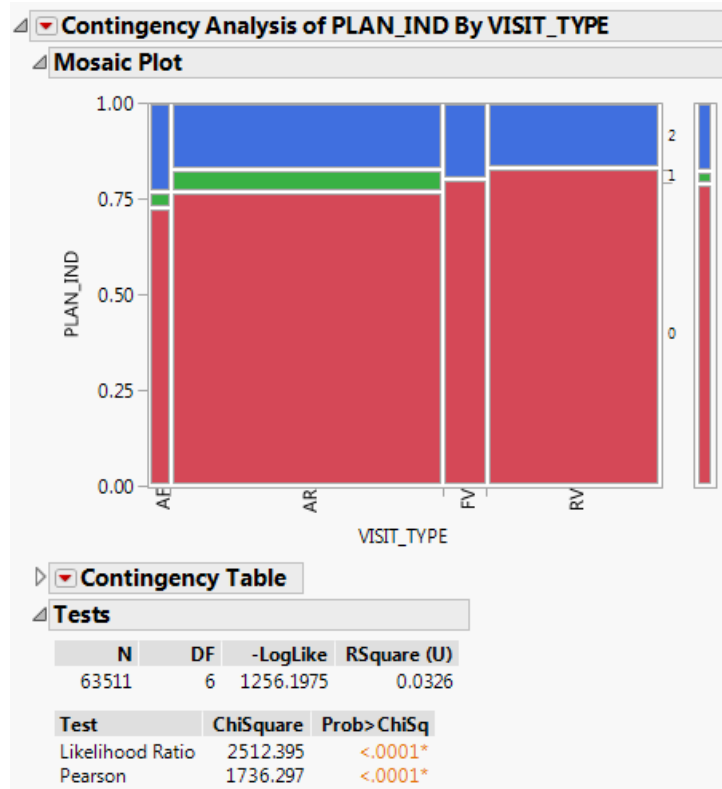
Gender



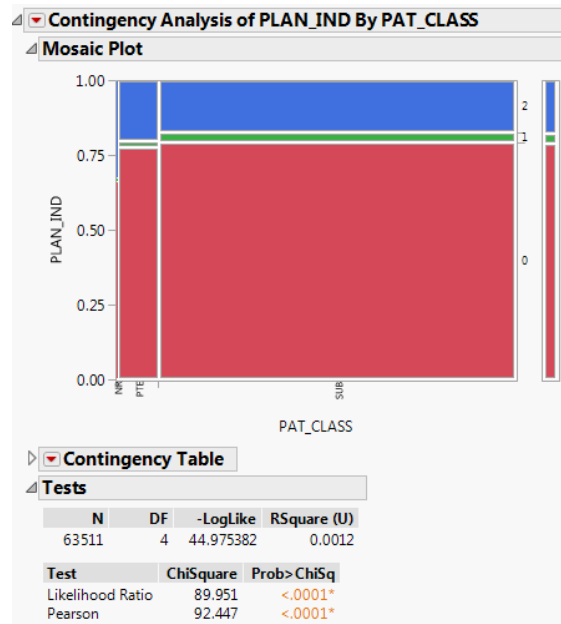
Reference Type



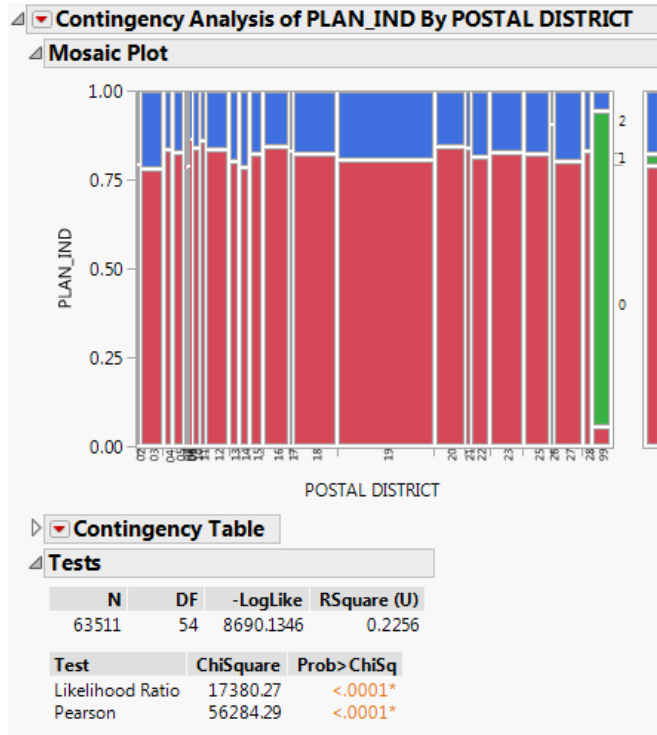
Visit Type



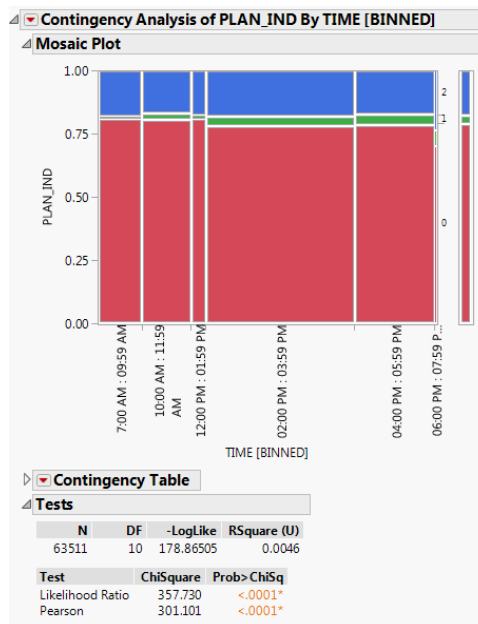
Patient Class



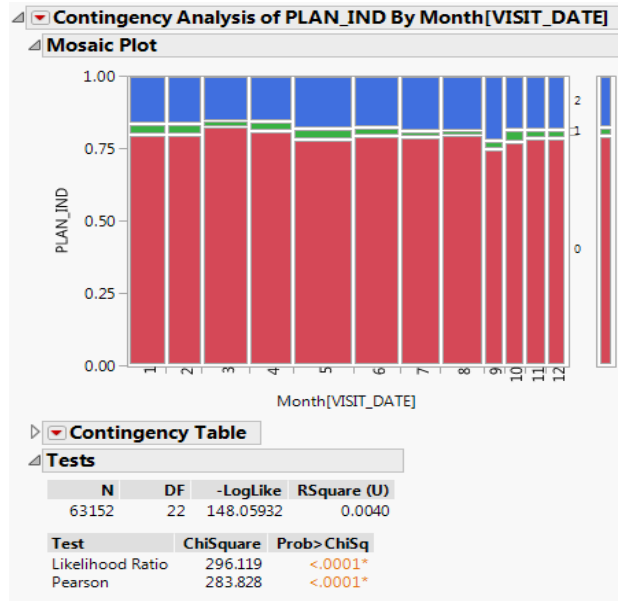
Postal District



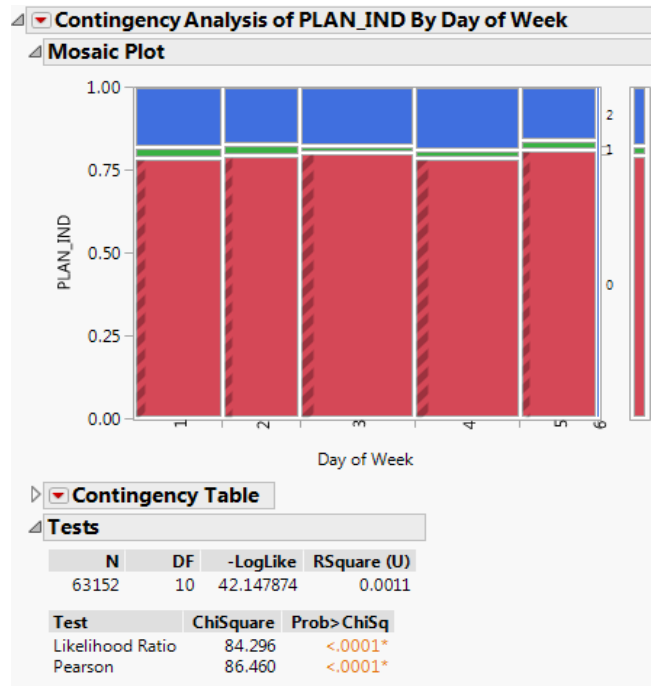
Time



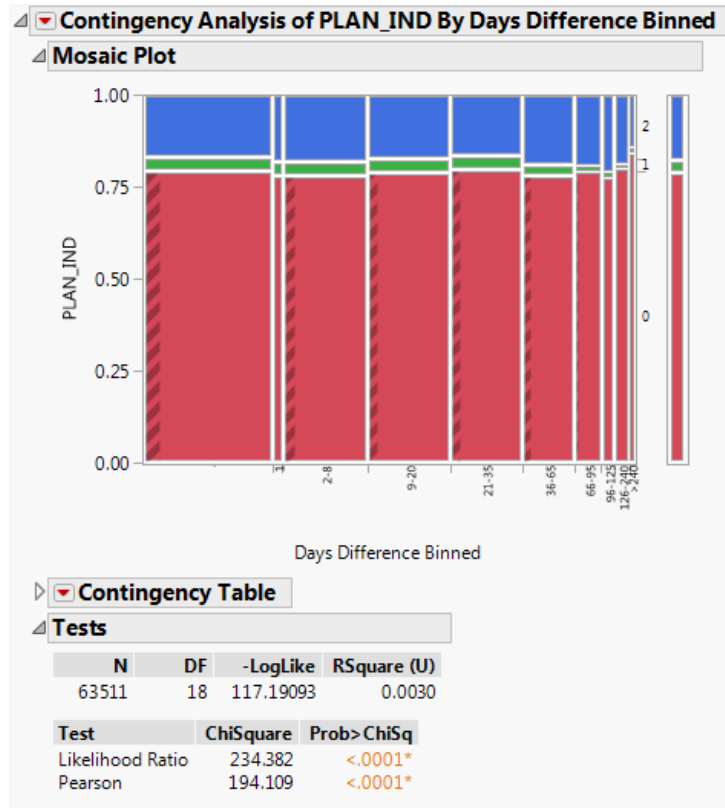
Month



Day



Appointment Age



Appendix 2.0 Logistic regression (Per episode for doctors) Parameter estimates

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.5514371	0.1684123	10.72	0.0011*
Race[C]	-0.6095268	0.0772704	62.22	<.0001*
Race[E]	0.18055662	0.1875226	0.93	0.3356
Race[I]	0.02030229	0.0928302	0.05	0.8269
Race[M]	-0.0390882	0.0848346	0.21	0.6450
Race[O]	-0.0749086	0.0954159	0.62	0.4324
Nationality[MY]	-0.0500346	0.1287185	0.15	0.6975
Nationality[Other]	-0.1036821	0.0917342	1.28	0.2584
Gender[F]	0.0414853	0.020994	3.90	0.0481*
Age	-0.0106373	0.0058525	3.30	0.0691
Clinic_ID[WCCGC]	-0.1907347	0.0231494	67.89	<.0001*
Visit_Type[FV]	0.06670141	0.0327376	4.15	0.0416*
Pat_Class[NR]	0.84010102	0.1560907	28.97	<.0001*
Pat_Class[PTE]	-0.2852494	0.0875141	10.62	0.0011*
Month[2-1]	0.08512048	0.0982221	0.75	0.3862
Month[3-2]	-0.1008754	0.0921179	1.20	0.2735
Month[4-3]	-0.0818944	0.0888757	0.85	0.3568
Month[5-4]	0.34783636	0.0844606	16.96	<.0001*
Month[6-5]	-0.2171096	0.07921	7.51	0.0061*
Month[7-6]	0.20538852	0.0870587	5.57	0.0183*
Month[8-7]	0.01388258	0.0874547	0.03	0.8739
Month[9-8]	0.04796261	0.1022541	0.22	0.6390
Month[10-9]	0.0961011	0.1155231	0.69	0.4055
Month[11-10]	-0.2156446	0.1178424	3.35	0.0673
Month[12-11]	0.19272395	0.1174892	2.69	0.1009
Day[2-1]	-0.0126562	0.0633549	0.04	0.8417
Day[3-2]	-0.0102558	0.0590002	0.03	0.8620
Day[4-3]	0.02639218	0.0544896	0.23	0.6281
Day[5-4]	-0.2279391	0.0693256	10.81	0.0010*
Neighbour[Immediate Districts]	-0.0794653	0.0392697	4.09	0.0430*
Neighbour[Other Districts]	-0.0550175	0.0369348	2.22	0.1363
Neighbour[WCCGC's District]	0.11230132	0.0689429	2.65	0.1033
Distance from Clinic (KM) Binned[0 — 5]	-0.0564999	0.0648805	0.76	0.3838
Distance from Clinic (KM) Binned[5 — 10]	0.01653386	0.0464771	0.13	0.7220
Distance from Clinic (KM) Binned[10 — 15]	-0.0178717	0.0457775	0.15	0.6962
Distance from Clinic (KM) Binned[15 — 20]	-0.0614314	0.0526278	1.36	0.2431
Ref_Type2[Comm MH Service]	0.09460282	0.0687058	1.90	0.1685
Ref_Type2[Intra-Hosp]	0.11376083	0.0487885	5.44	0.0197*
Ref_Type2[NHG Poly & Hosp]	-0.1025067	0.047128	4.73	0.0296*
Ref_Type2[Others]	0.20104291	0.0562334	12.78	0.0004*
Ref_Type2[School]	-0.0022952	0.070752	0.00	0.9741
Ref_Type2[Self]	-0.1876462	0.0906518	4.28	0.0385*
Appointment Age (Binned)[.]	-0.2329322	0.052929	19.37	<.0001*
Appointment Age (Binned)[1]	-0.3143557	0.1720858	3.34	0.0677
Appointment Age (Binned)[2-8]	-0.062891	0.067219	0.88	0.3495
Appointment Age (Binned)[9-20]	-0.0641976	0.0629786	1.04	0.3080
Appointment Age (Binned)[21-35]	-0.1000377	0.0596706	2.81	0.0936
Appointment Age (Binned)[36-65]	0.15159111	0.0558094	7.38	0.0066*
Appointment Age (Binned)[66-95]	0.24326494	0.0615165	15.64	<.0001*
Appointment Age (Binned)[96-125]	0.21026664	0.0783974	7.19	0.0073*
Appointment Age (Binned)[126-240]	0.2144782	0.0737674	8.45	0.0036*

Appendix 3.0 Logistic regression (Per episode for Allied Health Professionals) Evaluation

Whole Model Test				
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	527.726	112	1055.451	<.0001*
Full	17127.983			
Reduced	17655.708			
RSquare (U)		0.0299		
AICc		34484.9		
BIC		35419.1		
Observations (or Sum Wgts)		26971		

Figure 27: Whole Model Test for Logistic Regression Model (Allied Health Professionals)

Since the above results showed that p value $<.0001$, there is sufficient statistical evidence to reject the null hypothesis. The logistic model is useful to explain the odds of no-show patient appointments. In other words, the overall model is significant at the 0.001 level according to the Model chi-square statistic.

Lack Of Fit			
Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	49110	16371.659	32743.32
Saturated	49222	756.323	Prob>ChiSq
Fitted	112	17127.983	1.0000

Figure 28: Lack of Fit for Logistic Regression Model (Allied Health Professionals)

The above figure shows that lack of fit chi-square is insignificant (Prob>Chisq = 1.0000) and supports the conclusion that there is little to be gained by introducing additional variables.

Effect Likelihood Ratio Tests				
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Race	10	10	330.71386	<.0001*
Nationality	4	4	6.37032146	0.1731
Ref_Type2	12	12	59.544509	<.0001*
SEX	2	2	1.43163432	0.4888
Age	2	2	5.61633146	0.0603
Clinic_ID	2	2	2.58002594	0.2753
Visit_Type	2	2	49.5610472	<.0001*
Pat_Class	4	4	3.09343539	0.5423
Visit_Time (Binned)	10	10	151.022338	<.0001*
Month	22	22	205.956307	<.0001*
Day	10	10	45.1138895	<.0001*
Neighbour	6	6	32.6826084	<.0001*
Distance from Clinic (KM) Binned	8	8	9.05722067	0.3375
Appointment Age (Binned)	18	18	67.5308827	<.0001*

Figure 29: LRT for Logistic Regression Model (Allied Health Professionals)

In this case, not all predictors are significant. Nationality, Gender, Age, Clinic ID, Patient Class, Distance from Clinic are insignificant while the other variables have p -value $<.0001$. The parameter estimates report gives more detailed information on which category is significant within each predictor variable (Refer to Appendix 3.0).

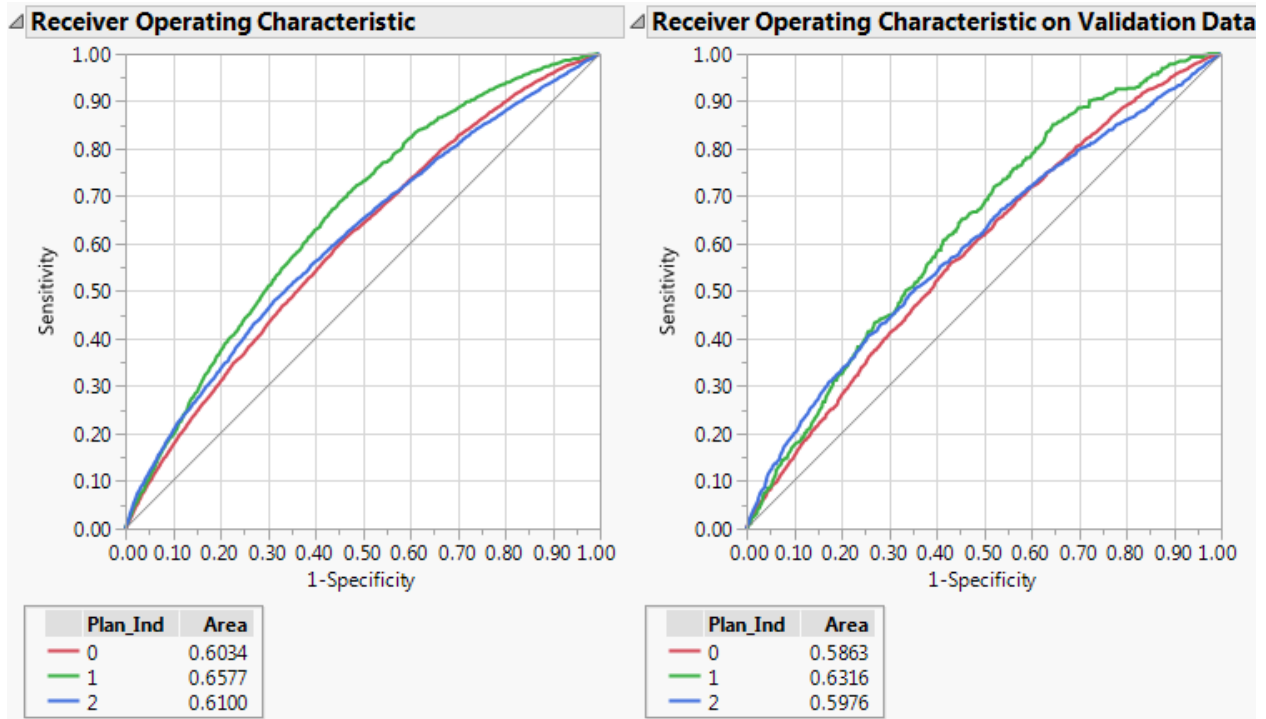


Figure 30: ROC Curve for Logistic Regression Model (Allied Health Professionals)

Similar to the logistic regression for doctors, the ROC indicates low distinguishability (not a very good model yet the model can be used).

Confusion Matrix							
Training			Validation				
Actual Plan_Ind	Predicted Count			Actual Plan_Ind	Predicted Count		
	0	1	2		0	1	2
0	20822	0	10	0	6865	0	3
1	1315	0	0	1	457	0	0
2	4813	0	11	2	1530	0	3

Figure 31: Confusion Matrix for Logistic Regression Model (Allied Health Professionals)

True Negative (Actual 0, Predict 0)	False Positive (Actual 0, Predict 2)
20,822	10
False Negative (Actual 2 & 1, Predict 0)	True Positive (Actual 2, Predict 2)
6,128	11

Table 12: Contingency Table for Logistic Regression Model (Allied Health Professionals)

The misclassification rate for the logistic regression model (allied health professionals) is 22.76%. Therefore, the model predicts 77.24% of the patient appointments attendance correctly. However, the model is only able to predict 0.23% of the no-show appointments. Like the logistic regression for doctors, we need to impute a new cut-off rate to gauge the probability of no-show appointments better.

		Plan_Ind			
		0	1	2	Total
Most Likely Plan_Ind	Count	8696	499	1346	10541
	Total %	32.24	1.85	4.99	39.08
	Col %	41.74	37.95	27.90	
	Row %	82.50	4.73	12.77	
[0.15]	Count	12136	816	3478	16430
	Total %	45.00	3.03	12.90	60.92
	Col %	58.26	62.05	72.10	
	Row %	73.86	4.97	21.17	
Total	20832	1315	4824	26971	
	77.24	4.88	17.89		

Figure 32: Contingency Table (15%) for Logistic Regression Model (Allied Health Professionals)

True Negative (Actual 0, Predict 0)	False Positive (Actual 0, Predict 2)
8,696	12,136
False Negative (Actual 2 & 1, Predict 0)	True Positive (Actual 2, Predict 2)
2,661	3,478

Table 13: Contingency Table (15%) for Logistic Regression Model (Allied Health Professionals)

The misclassification rate for 15% cut-off rate is 54.86%. Therefore, the model predicts 45.14% of the patient appointments attendance correctly. Based on true positives, the model is now able to predict 72.10% of the no-show patient appointments attendance correctly. In order to decide on the optimal cut-off rate, we repeated the computation of true positives for cut-off rate of 10%, 16%, 17%, 18%, 19%, 20%.

Term	Estimate	Std Error	ChiSquare	Prob> ChiSq
Intercept	0.75909969	0.227333	11.15	0.0008*
Race[C]	0.74371458	0.105969	49.26	<0.0001*
Race[E]	-0.1242041	0.1927232	0.42	0.5193
Race[I]	-0.0718663	0.1154356	0.39	0.5336
Race[M]	0.08363209	0.1110391	0.57	0.4503
Race[O]	0.37230913	0.1155918	10.37	0.0013*
Nationality[MV]	-0.0764423	0.1216331	0.39	0.5297
Nationality[Other]	0.13289817	0.0878648	2.29	0.1304
Ref_Type2[Comm MH Service]	-0.0694962	0.0589597	1.39	0.2385
Ref_Type2[Intra-Hosp]	-0.0472605	0.0432116	1.20	0.2741
Ref_Type2[NHG Poly & Hosp]	0.17854977	0.0418702	18.18	<0.0001*
Ref_Type2[Others]	-0.146526	0.0473269	9.59	0.0020*
Ref_Type2[School]	-0.0452059	0.0650857	0.48	0.4873
Ref_Type2[Self]	0.0097762	0.1125171	0.01	0.9308
SEX[F]	-0.0158668	0.0180529	0.77	0.3795
Age	-0.0106409	0.0052816	4.06	0.0439*
Clinic_ID[WCCGC]	0.01663876	0.0207448	0.64	0.4225
Visit_Type[AF]	-0.1990236	0.0285782	48.50	<0.0001*
Pat_Class[NR]	-0.0394446	0.3230228	0.01	0.9028
Pat_Class[PTE]	0.09732973	0.1734032	0.32	0.5746
Visit_Time (Binned)[7:00 AM : 09:59 AM]	0.12265253	0.0629327	3.80	0.0513
Visit_Time (Binned)[10:00 AM : 11:59 AM]	0.23319855	0.0616472	14.31	0.0002*
Visit_Time (Binned)[12:00 PM : 01:59 PM]	0.19673023	0.1122341	3.07	0.0796
Visit_Time (Binned)[02:00 PM : 03:59 PM]	-0.0014958	0.0519345	0.00	0.9770
Visit_Time (Binned)[04:00 PM : 05:59 PM]	-0.0090717	0.0556681	0.03	0.8705
Month[2-1]	-0.0690158	0.0800163	0.74	0.3884
Month[3-2]	0.13627988	0.0779098	3.06	0.0803
Month[4-3]	-0.0318825	0.0746181	0.18	0.6692
Month[5-4]	-0.1066312	0.0690174	2.39	0.1223
Month[6-5]	-0.0932067	0.067046	1.93	0.1645
Month[7-6]	-0.0117598	0.0703956	0.03	0.8673
Month[8-7]	0.14228034	0.0720138	3.90	0.0482*
Month[9-8]	-0.5020628	0.0843244	35.45	<0.0001*
Month[10-9]	0.50765552	0.1004065	25.56	<0.0001*
Month[11-10]	-0.0287731	0.1061813	0.07	0.7864
Month[12-11]	0.08520758	0.1107937	0.59	0.4419
Day[2-1]	0.13465174	0.0561883	5.74	0.0166*
Day[3-2]	0.01574703	0.0543579	0.08	0.7721
Day[4-3]	-0.1537799	0.047985	10.27	0.0014*
Day[5-4]	0.16202643	0.0505245	10.28	0.0013*
Day[6-5]	-18.805627	5301.0206	0.00	0.9972
Neighbour[Immediate Districts]	0.12055953	0.0313801	14.76	0.0001*

Neighbour[Other Districts]		0.14333547	0.0313108	20.96	<.0001*
Neighbour[WCCGC's District]		-0.2345849	0.0546859	18.40	<.0001*
Distance from Clinic (KM) Binned[0 — 5]		0.12360466	0.0579143	4.56	0.0328*
Distance from Clinic (KM) Binned[5 — 10]		0.02922671	0.0412677	0.50	0.4788
Distance from Clinic (KM) Binned[10 — 15]		0.01564602	0.0417778	0.14	0.7080
Distance from Clinic (KM) Binned[15 — 20]		-0.0500307	0.0480326	1.08	0.2976
Appointment Age (Binned)[.]		-0.0278264	0.0694074	0.16	0.6885
Appointment Age (Binned)[1]		-0.075503	0.109038	0.48	0.4887
Appointment Age (Binned)[2-8]		-0.0922829	0.0670729	1.89	0.1689
Appointment Age (Binned)[9-20]		0.05299671	0.0681272	0.61	0.4366
Appointment Age (Binned)[21-35]		0.08626024	0.0704198	1.50	0.2206
Appointment Age (Binned)[36-65]		0.00757347	0.0770247	0.01	0.9217
Appointment Age (Binned)[66-95]		-0.0218111	0.1062346	0.04	0.8373
Appointment Age (Binned)[96-125]		-0.353432	0.1465	5.82	0.0158*
Appointment Age (Binned)[126-240]		0.00323229	0.1656099	0.00	0.9844
Intercept	Unstable	-3.3476418	169.10981	0.00	0.9842
Race[C]	Unstable	2.64082256	169.10926	0.00	0.9875
Race[E]	Unstable	2.46368015	169.10954	0.00	0.9884
Race[I]	Unstable	2.33173718	169.10928	0.00	0.9890
Race[M]	Unstable	2.44160148	169.10927	0.00	0.9885
Race[O]	Unstable	2.24065086	169.10928	0.00	0.9894
Nationality[MV]		-0.1111471	0.2272903	0.24	0.6248
Nationality[Other]		0.26597742	0.1610191	2.73	0.0986
Ref_Type2[Comm MH Service]		-0.3475228	0.1295549	7.20	0.0073*
Ref_Type2[Intra-Hosp]		0.09111432	0.0838866	1.18	0.2774
Ref_Type2[NHG Poly & Hosp]		0.20949294	0.0792322	6.99	0.0082*
Ref_Type2[Others]		0.12236272	0.0890173	1.89	0.1693
Ref_Type2[School]		-0.1366041	0.132435	1.06	0.3023
Ref_Type2[Self]		0.06809807	0.218283	0.10	0.7551
SEX[F]		-0.0385549	0.034994	1.21	0.2706
Age		-0.0202659	0.10102599	3.90	0.0482*
Clinic_ID[WCCGC]		0.06417616	0.0402494	2.54	0.1108
Visit_Type[AF]		-0.2679262	0.0677656	15.63	<.0001*
Pat_Class[NR]		-0.1456676	0.7504518	0.04	0.8461
Pat_Class[PTE]		0.04538632	0.3949476	0.01	0.9085
Visit_Time (Binned)[7:00 AM : 09:59 AM]		-0.9317573	0.1525729	37.30	<.0001*
Visit_Time (Binned)[10:00 AM : 11:59 AM]		0.02451778	0.1206957	0.04	0.8390
Visit_Time (Binned)[12:00 PM : 01:59 PM]		0.39442665	0.2028589	3.78	0.0519
Visit_Time (Binned)[02:00 PM : 03:59 PM]		0.10973654	0.0988219	1.23	0.2668
Visit_Time (Binned)[04:00 PM : 05:59 PM]		0.3633056	0.1036881	12.28	0.0005*
Month[2-1]		-0.2785955	0.1407041	3.92	0.0477*
Month[3-2]		-0.4340786	0.1553813	7.80	0.0052*
Month[4-3]		0.53593348	0.1486865	12.99	0.0003*
Month[5-4]		0.10333748	0.1211084	0.73	0.3935
Month[6-5]		-0.4575922	0.126691	13.05	0.0003*
Month[7-6]		-0.4498644	0.1537862	8.56	0.0034*
Month[8-7]		-0.1907125	0.1787565	1.14	0.2860
Month[9-8]		0.28421972	0.197719	2.07	0.1506
Month[10-9]		0.82137836	0.1921832	18.27	<.0001*
Month[11-10]		-0.4428764	0.1960834	5.10	0.0239*
Month[12-11]		0.17307616	0.2128131	0.66	0.4161
Day[2-1]		-0.1266299	0.1021357	1.54	0.2150
Day[3-2]		-0.0842607	0.1034276	0.66	0.4153
Day[4-3]		-0.1349577	0.0978735	1.90	0.1679
Day[5-4]		0.1704907	0.100977	2.85	0.0913
Day[6-5]	Unstable	-14.767392	5301.0206	0.00	0.9978
Neighbour[Immediate Districts]		0.12139026	0.0628802	3.73	0.0535
Neighbour[Other Districts]		0.19086744	0.0621194	9.44	0.0021*
Neighbour[WCCGC's District]		-0.2886946	0.1165374	6.14	0.0132*
Distance from Clinic (KM) Binned[0 — 5]		-0.0503923	0.1144405	0.19	0.6597
Distance from Clinic (KM) Binned[5 — 10]		0.02247618	0.0771442	0.08	0.7708
Distance from Clinic (KM) Binned[10 — 15]		0.04418185	0.0783868	0.32	0.5730
Distance from Clinic (KM) Binned[15 — 20]		0.01896579	0.0904177	0.04	0.8339
Appointment Age (Binned)[.]		-0.5109175	0.1200492	18.11	<.0001*
Appointment Age (Binned)[1]		-0.0499815	0.201688	0.06	0.8043
Appointment Age (Binned)[2-8]		-0.274909	0.1120493	6.02	0.0141*
Appointment Age (Binned)[9-20]		0.02302759	0.1115897	0.04	0.8365
Appointment Age (Binned)[21-35]		0.08540679	0.1155157	0.55	0.4597
Appointment Age (Binned)[36-65]		0.12734918	0.1288664	0.98	0.3230
Appointment Age (Binned)[66-95]		0.09849101	0.1862744	0.28	0.5970
Appointment Age (Binned)[96-125]		-0.2169129	0.2842302	0.58	0.4454
Appointment Age (Binned)[126-240]		-0.1081779	0.3208844	0.11	0.7360

Appendix 4.0 Decision Tree Model (Per episode for Doctors) Column Contribution

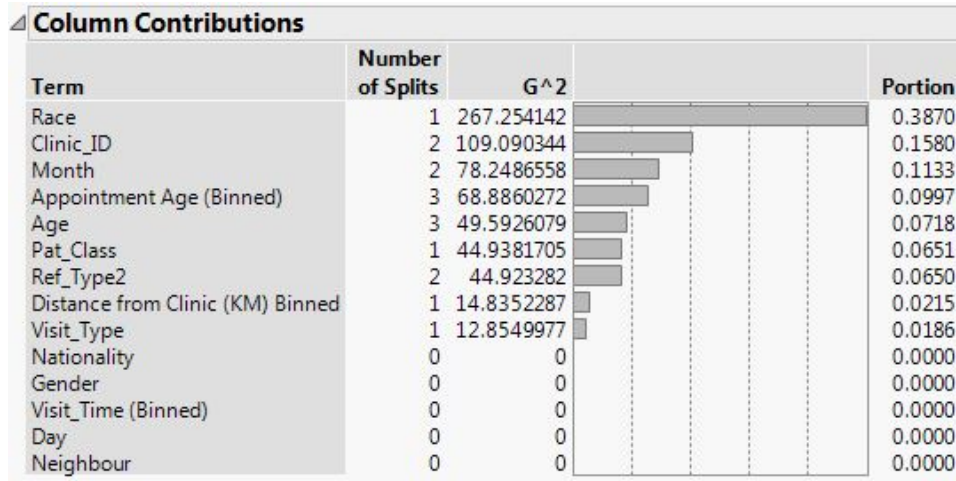


Figure 33: Column Contribution (Doctors)

Appendix 5.0 Decision Tree Model (Per episode for Allied Health Professionals) Evaluation

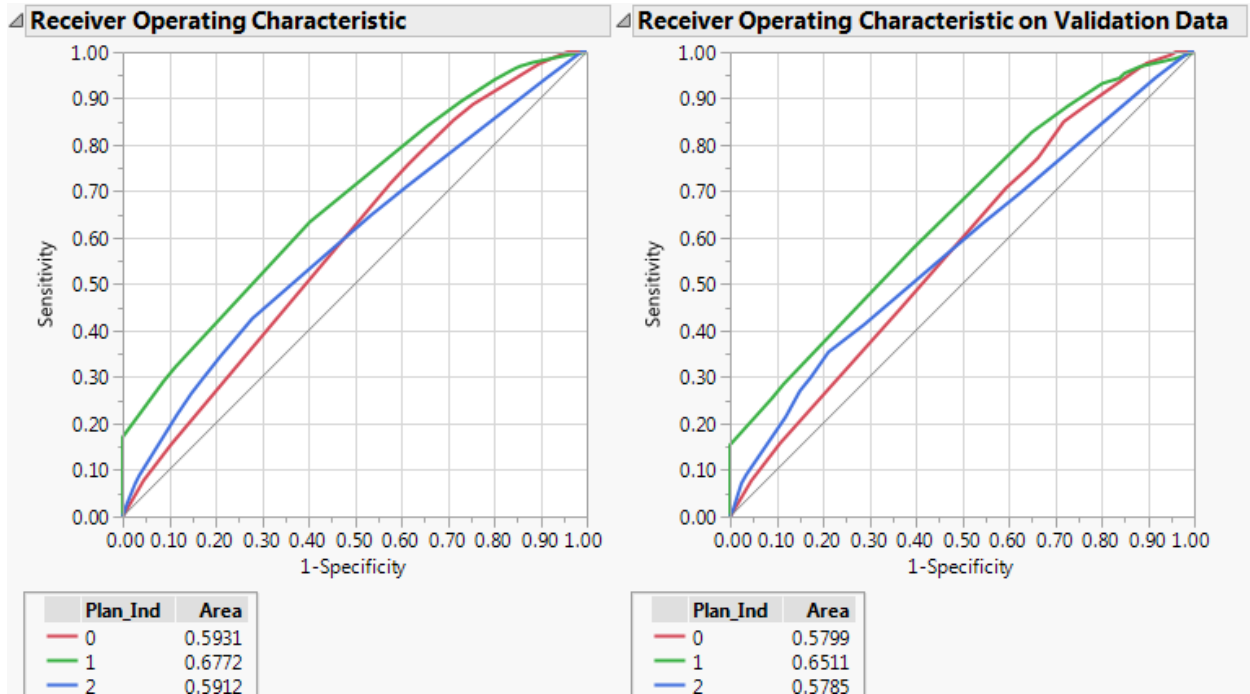


Figure 34: ROC Curve for Decision Tree Model (Allied Health Professionals)

The ROC curve for the decision tree model indicates a low distinguish ability (not a very good model yet the model can be used) in identifying no-show appointments from the model. The decision tree model for psychologists has a lower distinguishing power than that of the logistic regression model for psychologists.

Fit Details			
Measure	Training	Validation	Definition
Entropy RSquare	0.0627	0.0546	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.1098	0.0959	$(1 - (L(0)/L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
Mean -Log p	0.6345	0.6391	$\sum -\text{Log}(p[j])/n$
RMSE	0.4459	0.4461	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.3589	0.3590	$\sum y[j] - p[j] /n$
Misclassification Rate	0.2254	0.2233	$\sum (p[j] \neq p\text{Max})/n$
N	27629	9068	n

Confusion Matrix							
Actual Plan_Ind	Training			Actual Plan_Ind	Validation		
	Predicted Count 0	1	2		Predicted Count 0	1	2
0	21127	0	0	0	6958	0	0
1	1326	274	0	1	465	85	0
2	4902	0	0	2	1560	0	0

Figure 35: Fit Details & Confusion Matrix for Decision Tree Model (Allied Health Professionals)

The misclassification rate for the decision tree is 22.54%. Therefore, the model predicts 77.46% of the patient appointments attendance correctly. Based on the true positives, the decision tree model predicts 0% of the no-show appointments. Thus, we need to impute a new cut-off rate to gauge the probability of no-show appointments better.

Contingency Table					
[0.15] Most Likely Plan_Ind (Decision Tree)	Count	Plan_Ind			Total
		0	1	2	
0	9719	888	1728	12335	
	35.18	3.21	6.25	44.65	
	46.00	55.50	35.25		
	78.79	7.20	14.01		
2	11408	712	3174	15294	
	41.29	2.58	11.49	55.35	
	54.00	44.50	64.75		
	74.59	4.66	20.75		
Total	21127	1600	4902	27629	
	76.47	5.79	17.74		

Figure 36: Contingency Table (15%) for Decision Tree Model (Allied Health Professionals)

True Negative (Actual 0, Predict 0)	False Positive (Actual 0, Predict 2)
9,719	11,408
False Negative (Actual 2 & 1, Predict 0)	True Positive (Actual 2, Predict 2)
3,328	3,174

Table 14: Contingency Table (15%) for Decision Tree Model (Allied Health Professionals)

The misclassification rate for 15% cut-off rate is 53.21%. Therefore, the model predicts 46.79% of the patient appointments attendance correctly. Based on true positives, the model is now able to predict 64.74% of the no-show patient appointments attendance correctly. In order to decide on the optimal cut-off rate, we repeated the computation of true positives for cut-off rate of 10%, 16%, 17%, 18%, 19%, 20%.

Appendix 6.0 Decision Tree Model (Per episode for Allied Health Professionals) Column Contribution

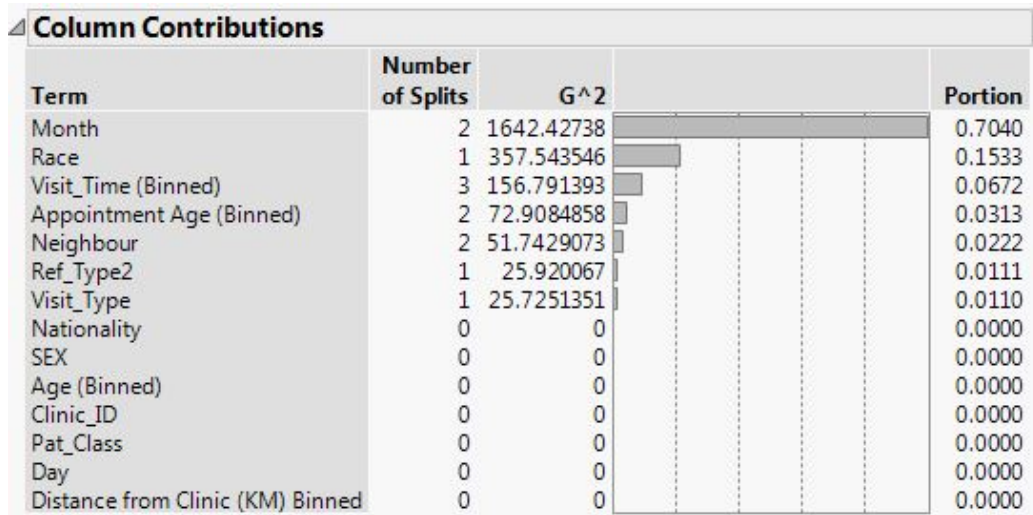


Figure 33: Column Contribution (Allied Health Professionals)