ANLY482: Analytics Practicum

Analysis of No-Show Appointments for Hospital X

Prof. Kam Tin Seong

Group 18 | Team ZAN

Loh Yan Zoey, Mirania Aishwarya Agarwal, Nasrullah Bin Khairullah

Presentation Outline

1. INTRODUCTION

> No-show appointment is defined as when a patient does not attend for a scheduled clinic appointment or cancels it with such minimal lead time that the slot cannot be filled

[Huang & Hanauer, 2014]

Study Context

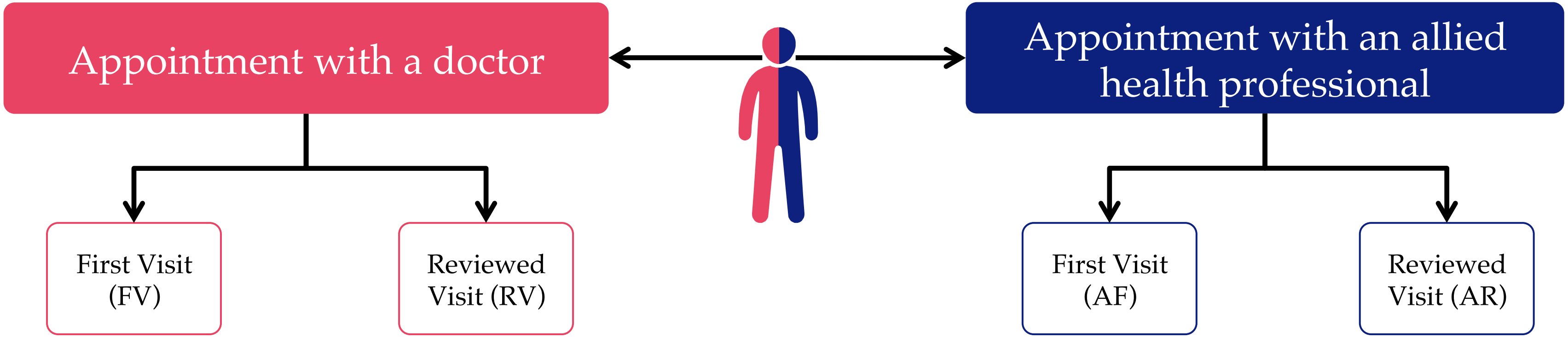
<u>Problems of No-Show Appointments</u>

- Patients missed an opportunity for a medical consultation
- Disruption of clinics' operations
- Decreased access to care for other patients

Project Sponsor: Hospital X

Project Background

- No-show appointment rate: 21% for first visits
- No-show appointment rate: 19% for review visits



Project Objective

To identify the significant factors that relate to no-show appointments and predict the no-show outcome from patients' appointments

Our Data

- 77,205 outpatient records across two clinics of Hospital X (2015-2016)
- Records are processed by frontline staff
- Patients are below 25 years old
- Most variables are categorical

2. LITERATURE REVIEW

Literature Review

<u>Similarity</u>

- Demographic variables (Age, gender, etc.)
- Appointment variables (Time, day, etc.)

<u>Differences</u>

- Financial information
- Appointment age*
- Distance of patients' residence to location of clinic*
- Appointment reminders

Literature Review

<u>Ma, Seemanta, Wu and Ng (2014)</u>

- Developed logistic regression & recursive partitioning models for 3 clinics in Singapore
- Included financial debt and reminder responses as predictor variables
- Results showed variations in significant predictor variables for no-show appointments among the 3 clinics

3. METHODOLOGY

Methodology

Original Data → Data Cleaning & Preparation → Analytical Sandboxes Preparation → Model Building → Results

4. DATA PREPARATION

Identify Missing Data

- Used missing data pattern in JMP Pro
- Cross referenced all records of a patient
- Filled in the missing value for the same patient

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

Rectifying Duplications & Discrepancies

- Used recode function to standardize names
- Rectified inconsistency in the recording of gender and nationality of patients

28743	V9_3382	C	M
28744	V9_3382	C	M
28745	V9_3382	C	U
28746	V9_3382	C	M
28747	V9_3382	C	M
28748	V9_3382	C	M
28749	V9_3382	C	M

Data Binning

Age

- 0 to 5 years old
- 6 to 10 years old
- 11 to 15 years old
- 16 to 20 years old
- 21 to 25 years old

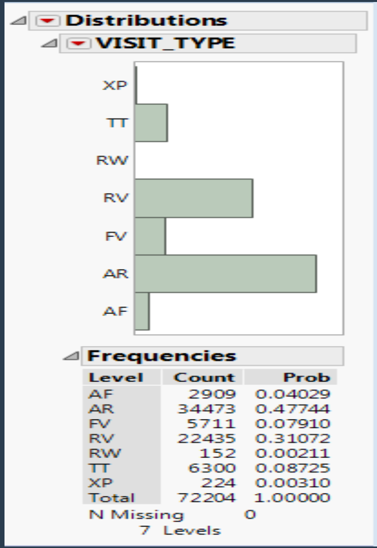
Appointment Timing

- 07:00am to 09:59am
- 10:00am to 11:59am
- 12:00pm to 01:59pm
- 02:00pm to 03:59pm
- 04:00pm to 05:59pm
- 06:00pm to 07:59pm

Variable & Dimension Reduction

- Removed irrelevant variables such as ‘RW’, ’TT’, ‘XP’
- Combined insignificant values within variables

- E.g. ‘Others’ & ‘None’ for Race



Level	Count	Prob
AF	2909	0.04029
AR	34473	0.47744
FV	5711	0.07910
RV	22435	0.31072
RW	152	0.00211
TT	6300	0.08725
XP	224	0.00310
Total	72204	1.00000

N Missing 0
7 Levels

New Variables Derived

<u>Appointment Age</u>

- Sort the data by patient ID and visit date
- Calculate the lead time between a patient's previous scheduled appointment and the next scheduled appointment.

New Variables Derived

<u>Clinic Switch</u>

- Filter the data to obtain patients who have visited both clinics at least once
- Sort data by patient and visit date
- A clinic switch (denoted as 1) occurs whenever the next scheduled appointment's clinic is different from the previous appointment's clinic

New Variables Derived

<u>Distance of Patient's Residence from location of each clinic</u>

- Update patients' postal codes
- Generate longitudes & latitudes from postal codes
- Convert WGS 84 coordinates to SVY21
- Formulae distances of patients' residence to each clinic

Data Preparation Process

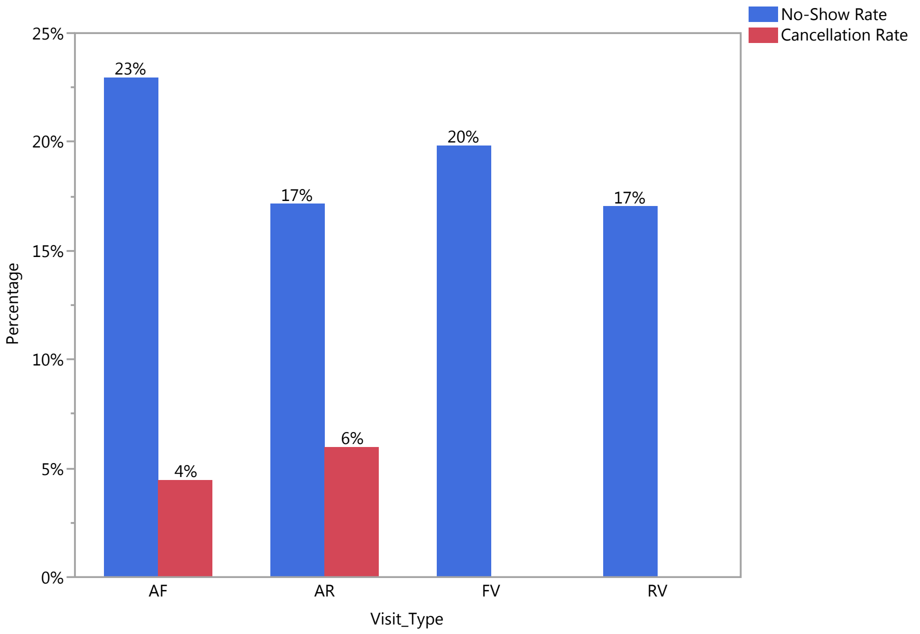
Identify Missing Data → Rectify Duplications & Discrepancies → Aggregate certain variables → Reduce variables' dimension → Derive new variables

Post-Data Preparation Process: 63,511 records left (82% of data retained)

5. FINDINGS

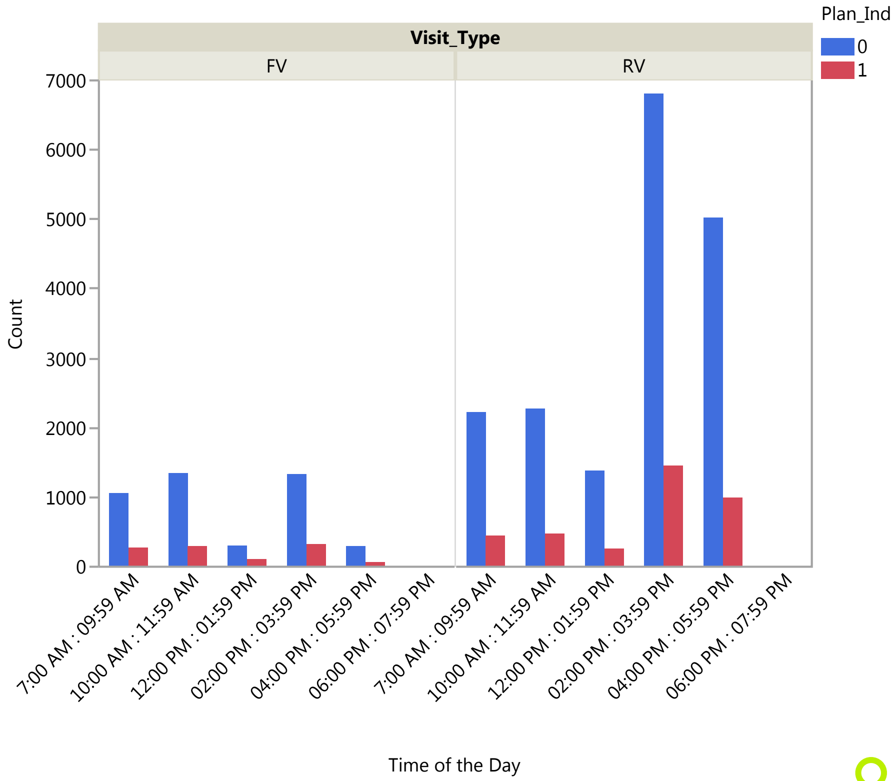
Visit Type Analysis

- Higher no-show rate for first visits than reviewed visits
- No cancellation rate for appointments under doctors



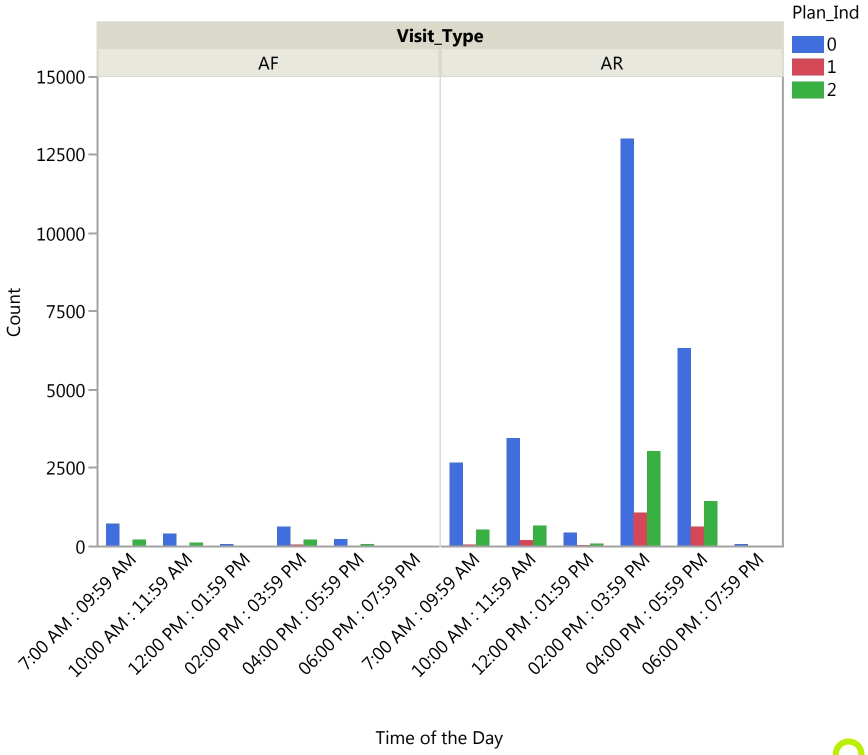
Visit Types (Doctor) Analysis

- More reviewed appointments scheduled than first appointments
- Stronger preference for late afternoon schedule



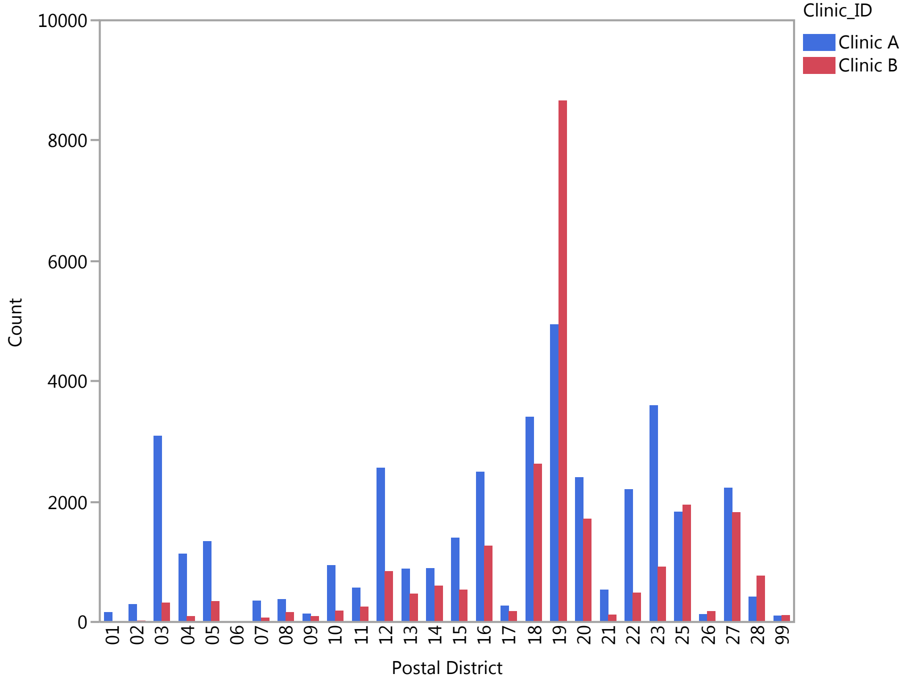
Visit Types (Allied Health Professional) Analysis

- Shared similar characteristics to appointments under doctors



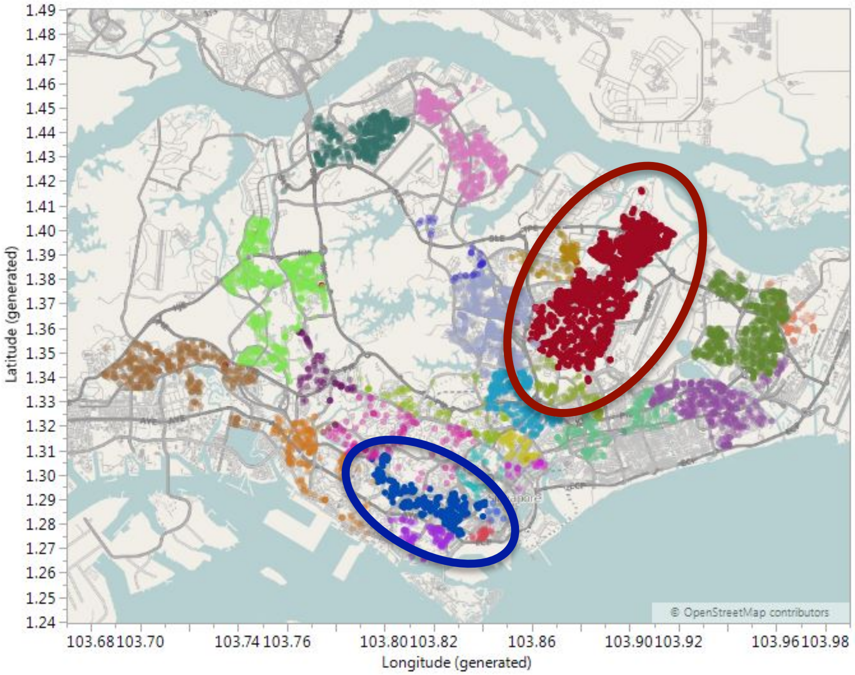
Distribution of Patient Analysis

- District 19 has the highest number of patients visiting both clinics
- In District 19, Clinic B has a significant portion of patients than Clinic A



Distribution of Patient Analysis

- Clinic A is located in district 3
- Clinic B is located in district 19
- A high density of patients living in Clinic B's district



6. ANALYTICAL SANDBOX

Analytical Sandbox

Data	Plan_IND	Models
Per episode for Doctors	0- Attended, 1- No-show	Logistic Regression & Decision Tree
Per episode for Allied health professionals	0- Attended, 1- Cancelled, 2- No-show	Multinomial Logistic Regression & Decision Tree
Per patients	0- Attended, 1- Cancelled, 2- No-show	Multiple Linear Regression

7. LOGISTIC REGRESSION MODEL

Logistic Regression

- The dependent variable, Plan IND has categorical responses
- Logistic regression deals with categorical response variable by using a logarithmic transformation on the response variable

Dealing with Multicollinearity

- Logistic regression is sensitive to high correlation among independent variables
- Performed chi-square tests
- Ensure correlation is p-value ≤ 0.05

Contingency Analysis of Plan_Ind By Appointment Age (Binned)

Mosaic Plot

Contingency Table

Tests

N	DF	-LogLike	RSquare (U)
63511	18	117.19093	0.0030

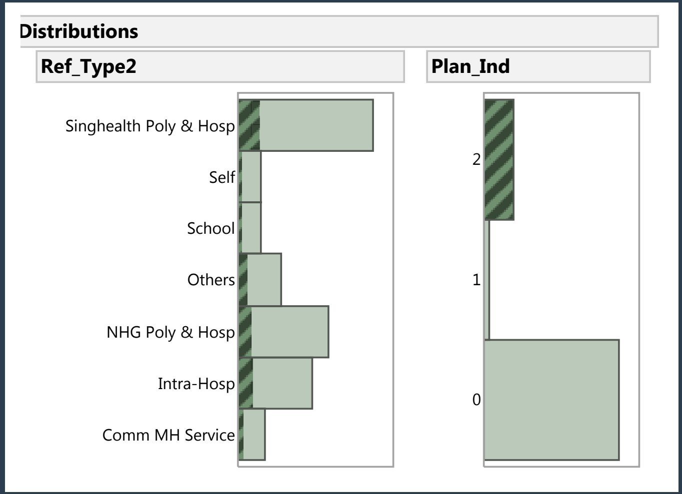
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	234.382	<.0001*
Pearson	194.109	<.0001*

Measures of Association

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

Dealing with Complete or Quasi-Complete Separation

- Occurs when a predictor variable is able to predict the response variable perfectly
- Make sure that the response variable is not a dichotomous version of another variable in the model



Logistic Regression Model Evaluation (Doctor)

H_0: The model is not useful

H_1: The model is useful

<u>Whole Model Test</u>

- The logistic model is useful in explaining the odds of appointments' attendance for doctor

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

Logistic Regression Model Evaluation (Doctor)

H_0: The model is adequate

H_1: The model is inadequate

<u>Lack of Fit Test</u>

- The logistic model is adequate in explaining the odds of appointments' attendance for doctors

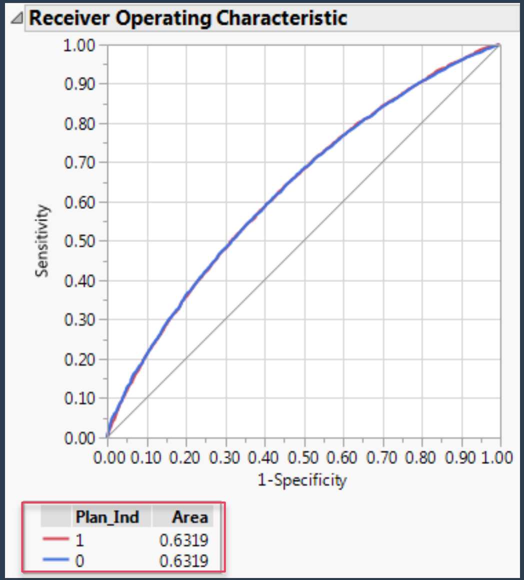
Lack Of Fit

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

Logistic Regression Model Evaluation (Doctor)

<u>ROC Curve</u>

- Indicates a low distinguish ability (not a very good model, yet the model can be used) in identifying appointments' attendance for doctors



Logistic Regression Model Evaluation (Doctor)

Confusion Matrix	
True Negatives	**False Positives**
16,285	10
False Negatives	**True Positives**
3,406	14

- The model is able to predict **82.67%** of appointments' attendance for doctors correctly
- However, it is only able to predict **0.41%** of no-show appointments

Logistic Regression Model Evaluation (Doctor)

- JMP Pro's prediction formula is based on the cut-off rate of 0.50
- The data has a significantly portion of attended appointments as compared to no-show appointments
- Need to compute a new cut-off rate to predict no-show appointments better

Logistic Regression Model Evaluation (Doctor)

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

Logistic Regression Model Evaluation (Doctor)

Effect Likelihood Ratio Test	
Parameters	**Prob>ChiSq**
Race	<.0001*
Nationality	0.0375*
Gender	0.0486*
Age	0.0692
Clinic ID	<.0001*
Visit Type	0.0418*

Logistic Regression Model Evaluation (Doctor)

Effect Likelihood Ratio Test	
Parameters	**Prob>ChiSq**
Patient Class	<.0001*
Month	<.0001*
Day	0.0087*
Neighbour	0.1472
Distance from Clinic	0.5743
Referral Type	<.0001*
Appointment Age	<.0001*

8. DECISION TREE MODEL

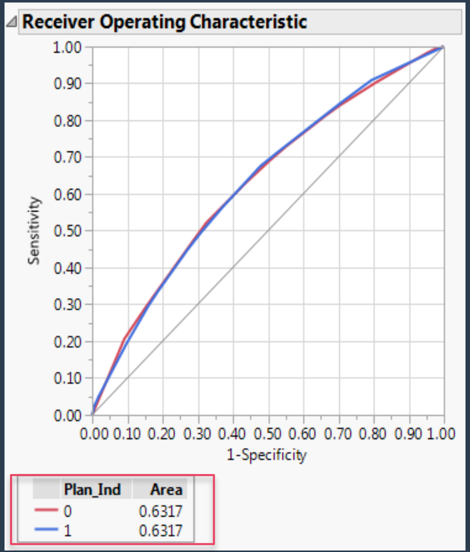
Decision Tree Model

- Predicts future observations based on decision rules that recursively splits independent variables into homogeneous zones
- Able to handle incomplete data
- Does not require any statistical assumptions regarding the data

Decision Tree Model Evaluation (Doctor)

<u>ROC Curve</u>

- Indicates a low distinguish ability (not a very good model, yet the model can be used) in identifying appointments' attendance for doctors



Decision Tree Model Evaluation (Doctor)

Confusion Matrix	
True Negatives	**False Positives**
16,559	0
False Negatives	**True Positives**
3527	0

- The model is able to predict **82.44%** of appointments' attendance for doctors correctly
- However, it is only able to predict **0%** of no-show appointments

Decision Tree Model Evaluation (Doctor)

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

Decision Tree Model Evaluation (Doctor)

Column Contribution		
Parameters	**G^2**	**Portion**
Race	267.254	0.3870
Clinic ID	109.090	0.1580
Month	78.249	0.1133
Appointment Age	68.886	0.0997
Age	49.593	0.0718

Decision Tree Model Evaluation (Doctor)

Column Contribution		
Parameters	**G^2**	**Portion**
Patient Class	44.938	0.0651
Referral Type	44.923	0.0650
Distance from Clinic	14.835	0.0215
Visit Type	12.855	0.0186

9. MODEL COMPARISON

Model Comparison (Doctor)

Predictive Performance Metrics (Based on Default JMP Pro Prediction Formula)		
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

Model Comparison (Doctor)

Predictive Performance Metrics (Based on Optimal Cut-off Rate)		
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%

Model Comparison (Doctor)

Common Significant Factors

- Race
- Clinic ID
- Patient Class
- Month
- Referral Type
- Appointment Age

10. CONCLUSION

Conclusion

- Current iteration of models can still be improved in terms of its explanatory & predictive ability
- Analysis may benefit from more than one year of data
- Other possible factors to consider are appointment reminders and number of people in the household of the patient

Thank You