Prediction of patients' show up for appointment with logistic regression model

Group member: Gan Luan, Liping Li,
Xindi Ruan, Xiao Liu


Doctor
Appointment
DATE
TIME
PICK UP
On Time
CALENDAR
APRIL
2015

Introduction

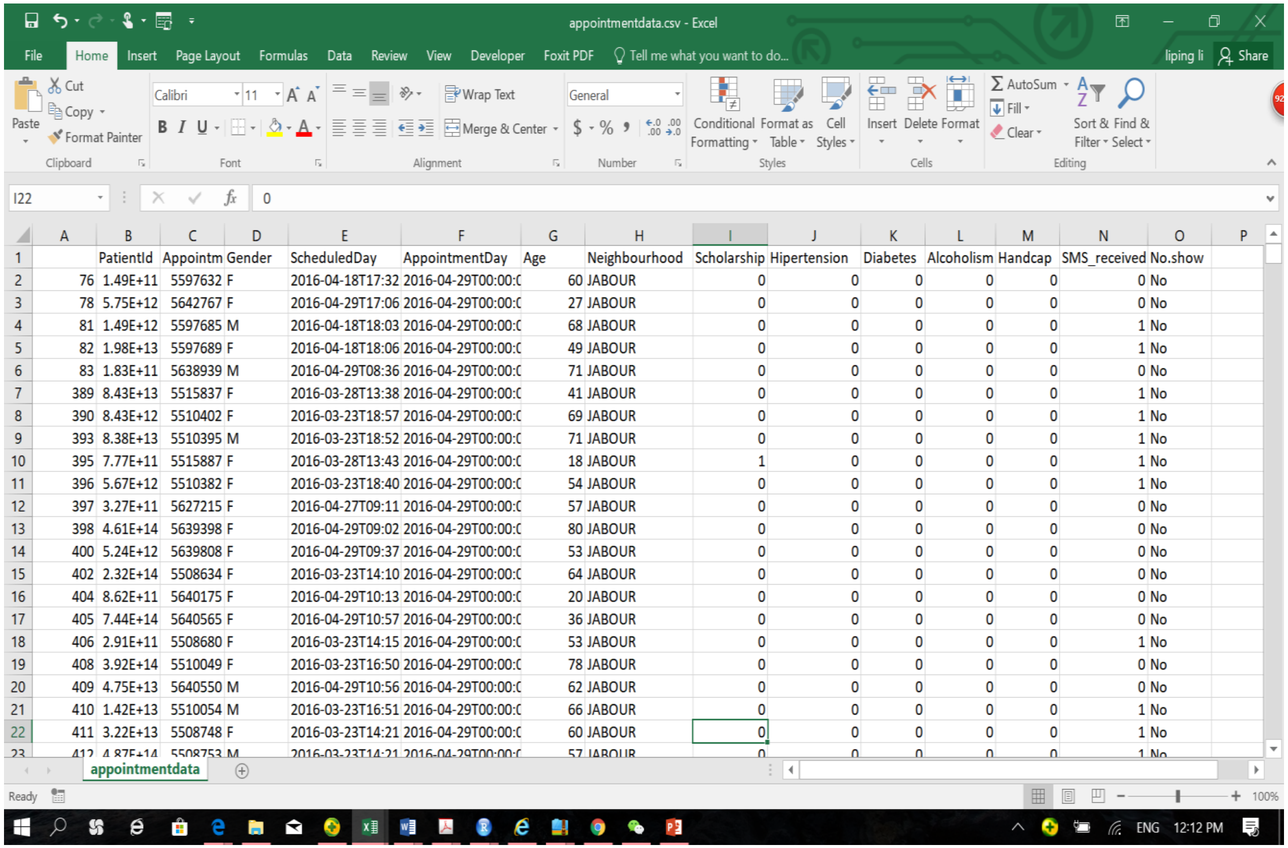
Problem:

- Around 20% patients never show up in their scheduled appointment. How can we predict patients’ show up?

Solution:

- Selecting one neighborhood- JABOUR to fit model from 110k medical appointments data set.
- Choosing 7 of 15 variables include one derived variable .
- Using Decision Tree, and Cross Validation to fit the model.

Variables

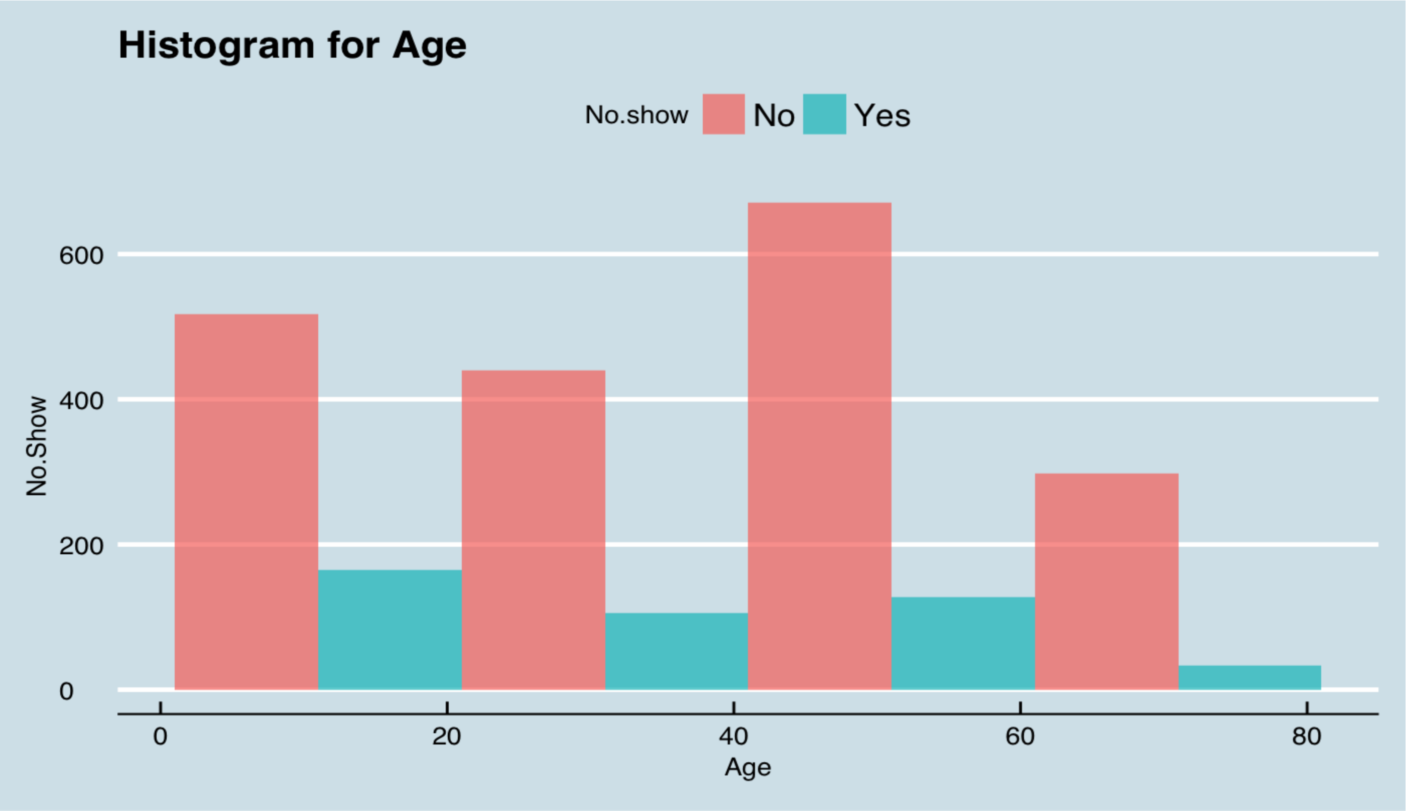


	PatientId	Appointm	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No.show
76	1.49E+11	5597632	F	2016-04-18T17:32	2016-04-29T00:00:0	60	JABOUR	0	0	0	0	0	0	No
78	5.75E+12	5642767	F	2016-04-29T17:06	2016-04-29T00:00:0	27	JABOUR	0	0	0	0	0	0	No
81	1.49E+12	5597685	M	2016-04-18T18:03	2016-04-29T00:00:0	68	JABOUR	0	0	0	0	0	1	No
82	1.98E+13	5597689	F	2016-04-18T18:06	2016-04-29T00:00:0	49	JABOUR	0	0	0	0	0	1	No
83	1.83E+11	5638939	M	2016-04-29T08:36	2016-04-29T00:00:0	71	JABOUR	0	0	0	0	0	0	No
389	8.43E+13	5515837	F	2016-03-28T13:38	2016-04-29T00:00:0	41	JABOUR	0	0	0	0	0	1	No
390	8.43E+12	5510402	F	2016-03-23T18:57	2016-04-29T00:00:0	69	JABOUR	0	0	0	0	0	1	No
393	8.38E+13	5510395	M	2016-03-23T18:52	2016-04-29T00:00:0	71	JABOUR	0	0	0	0	0	1	No
395	7.77E+11	5515887	F	2016-03-28T13:43	2016-04-29T00:00:0	18	JABOUR	1	0	0	0	0	1	No
396	5.67E+12	5510382	F	2016-03-23T18:40	2016-04-29T00:00:0	54	JABOUR	0	0	0	0	0	1	No
397	3.27E+11	5627215	F	2016-04-27T09:11	2016-04-29T00:00:0	57	JABOUR	0	0	0	0	0	0	No
398	4.61E+14	5639398	F	2016-04-29T09:02	2016-04-29T00:00:0	80	JABOUR	0	0	0	0	0	0	No
400	5.24E+12	5639808	F	2016-04-29T09:37	2016-04-29T00:00:0	53	JABOUR	0	0	0	0	0	0	No
402	2.32E+14	5508634	F	2016-03-23T14:10	2016-04-29T00:00:0	64	JABOUR	0	0	0	0	0	0	No
404	8.62E+11	5640175	F	2016-04-29T10:13	2016-04-29T00:00:0	20	JABOUR	0	0	0	0	0	0	No
405	7.44E+14	5640565	F	2016-04-29T10:57	2016-04-29T00:00:0	36	JABOUR	0	0	0	0	0	0	No
406	2.91E+11	5508680	F	2016-03-23T14:15	2016-04-29T00:00:0	53	JABOUR	0	0	0	0	0	1	No
408	3.92E+14	5510049	F	2016-03-23T16:50	2016-04-29T00:00:0	78	JABOUR	0	0	0	0	0	0	No
409	4.75E+13	5640550	M	2016-04-29T10:56	2016-04-29T00:00:0	62	JABOUR	0	0	0	0	0	0	No
410	1.42E+13	5510054	M	2016-03-23T16:51	2016-04-29T00:00:0	66	JABOUR	0	0	0	0	0	1	No
411	3.22E+13	5508748	F	2016-03-23T14:21	2016-04-29T00:00:0	60	JABOUR	0	0	0	0	0	1	No
412	4.87E+14	5508753	M	2016-03-23T14:21	2016-04-29T00:00:0	57	JABOUR	0	0	0	0	0	1	No

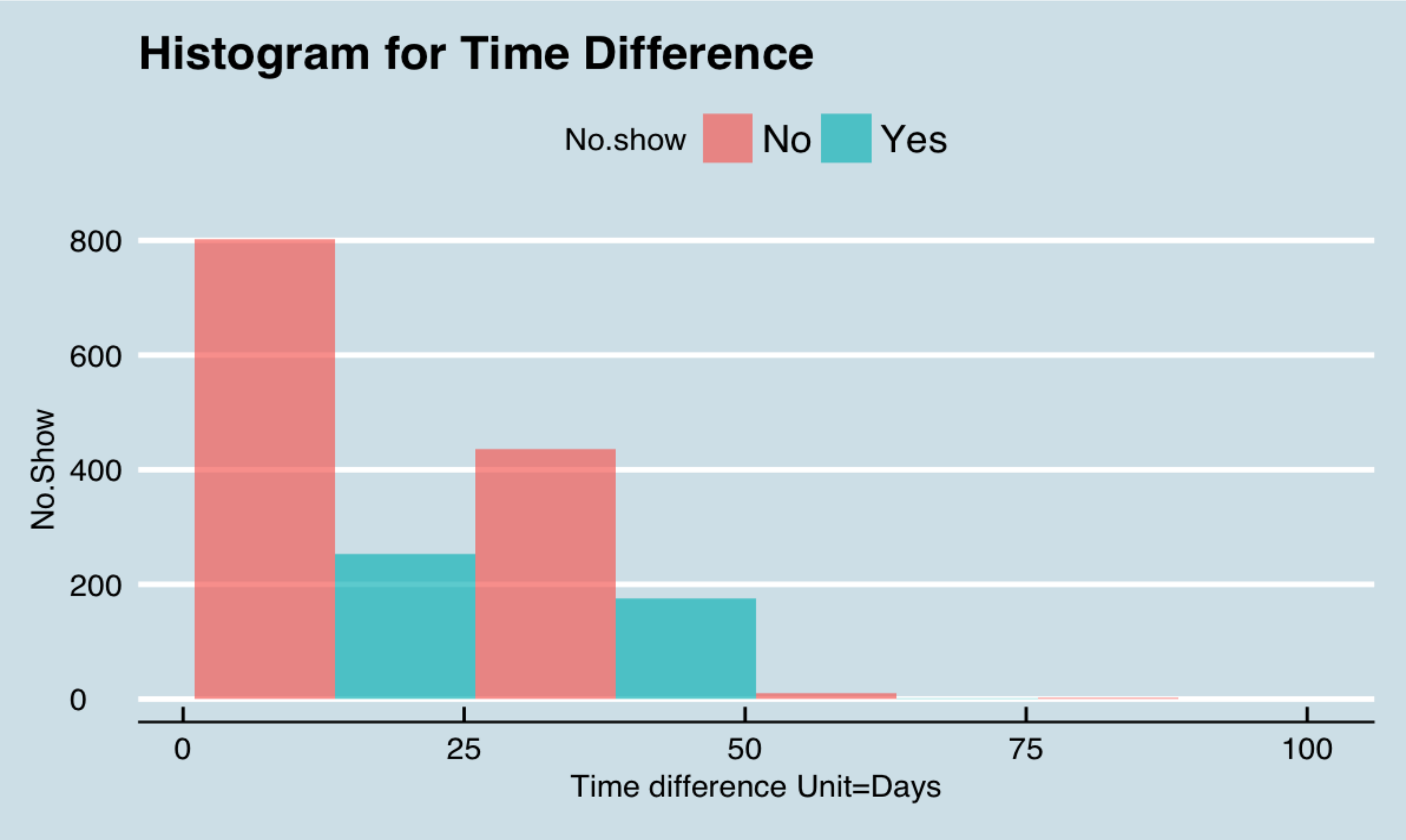
https://www.kaggle.com/joniarroba/noshowappointments

PatientID
AppointmentID
Gender
Scheduleday
Appointmentday
Age
Neighbourhood
Scholarship
Hypertension
Diabetes
Alcoholism
Handicap
SMS_received
Difftime
No-show

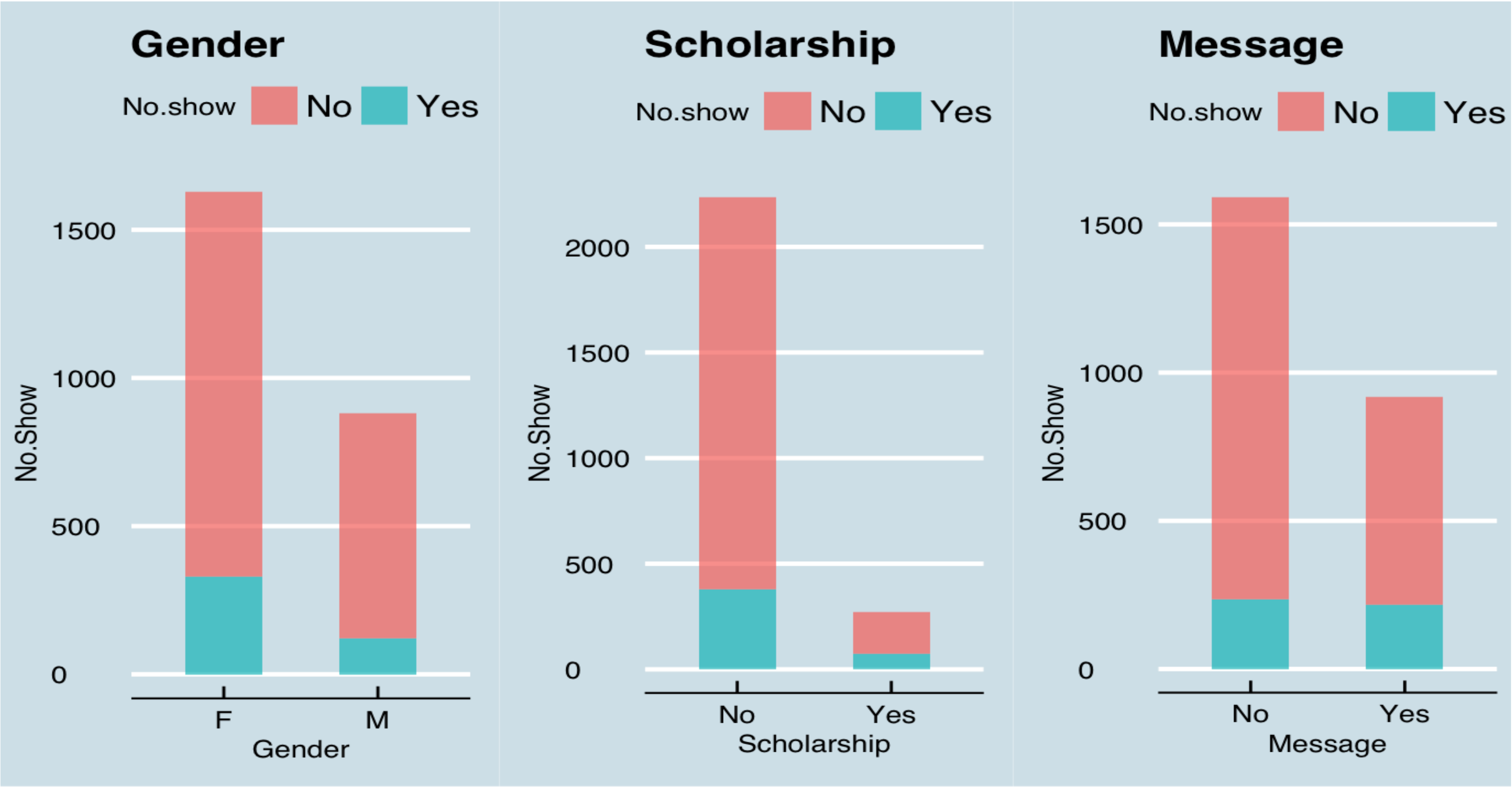
What do variables tell us?



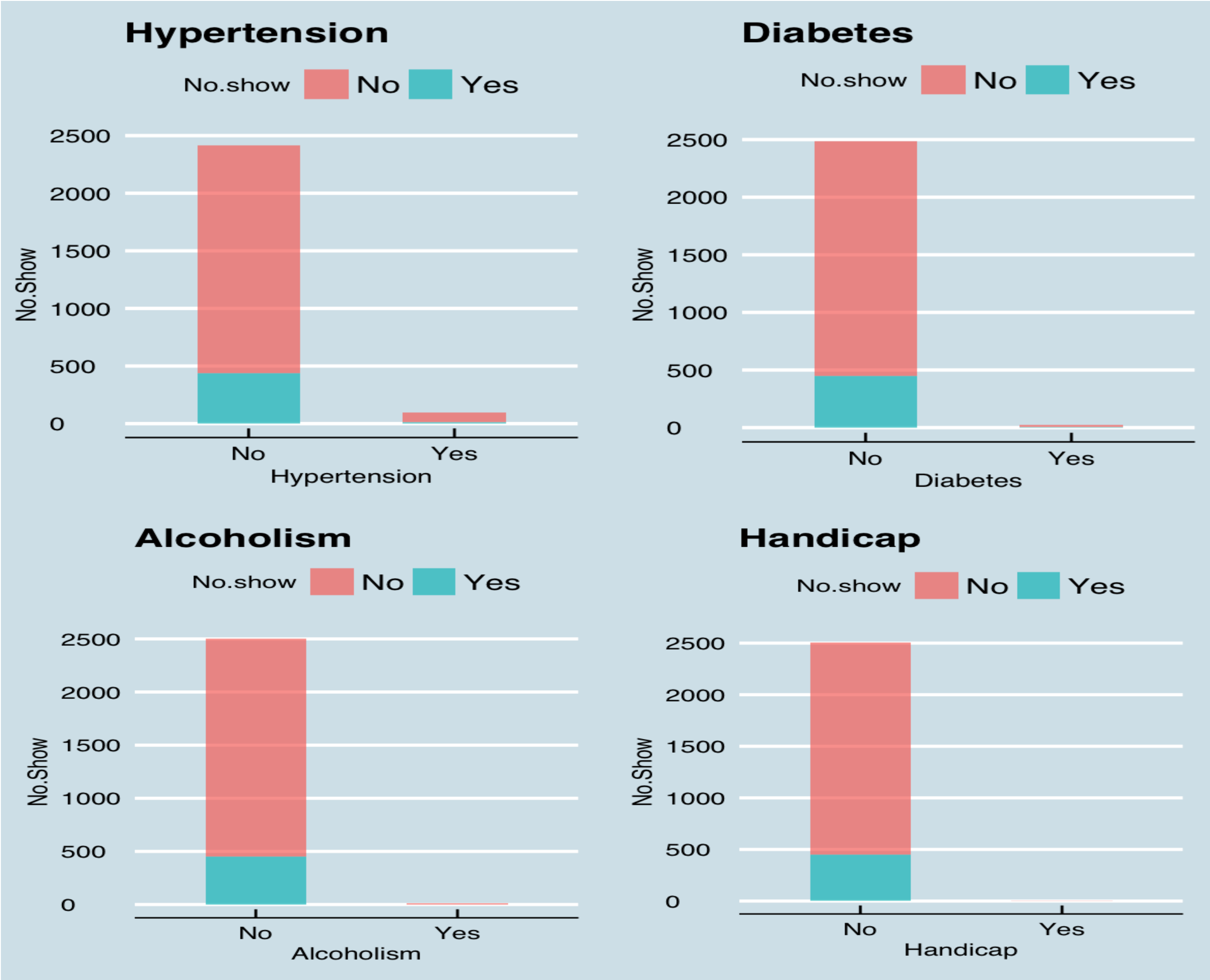
What do variables tell us?



Bar graphs of Categorical Variables



Bar graphs of Categorical Variables
Patient History Variables



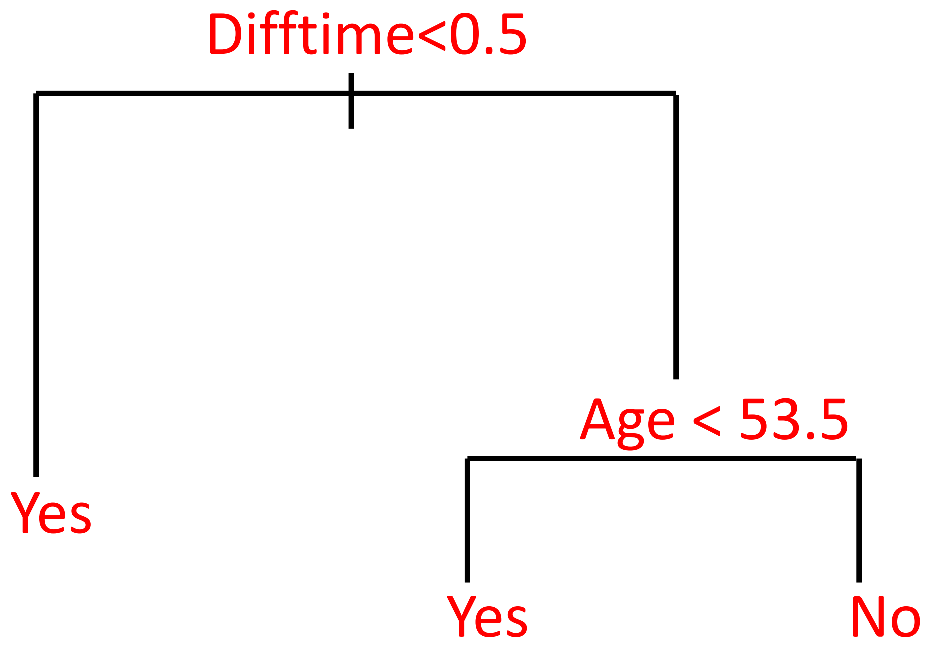
Decision tree

Why we choose decision tree

- Easier to explain and interpret
- More closely reflect the human decision-making
- Straightforward for qualitative variables
- Our model has several qualitative predictors, decision tree is a good fit.

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani , An Introduction to Statistical Learning with Applications in R, 1st ed. 2013, Corr. 7th printing 2017 Edition, Springer, 2013

Decision tree



Cross validation to choose the best model

All possible models

Logistic regression model was chosen since our response variable is binary.

The possible number of variables we can contain in our models are 1, 2, …, 7.

For each number of variables, we fit all the possible models.

For example, if we are allowed to contain 1 parameter, we can fit 7 models. If we are allowed to contain 2 parameters, we can fit $\binom{7}{2} = 21$ models.

Cross validation to choose the best model

Cross validation

For each model, we use cross validation to calculate prediction error.

Cross validation to choose the best model

Error calculation

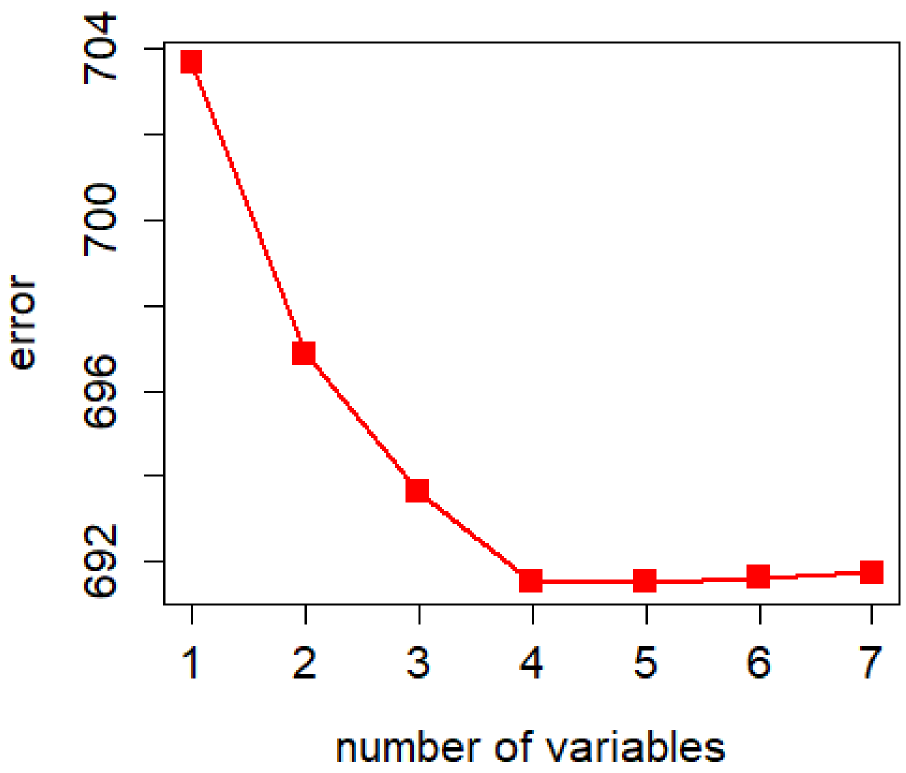
$$\sum_i e_i \,,$$

$$e_i = \begin{cases} 1 - predicted\ probablity, & real\ response = 'Yes' \\ predicted\ prbability - 0, & real\ response = 'No' \end{cases}$$

Response	Predicted prob	e_i
Yes	0.9	0.1
Yes	0.1	0.9
No	0.1	0.1
No	0.9	0.9

Cross validation to choose the best model

Lowest error vs number of variables



When 2 variables are allowed, variables are: Age and difftime.

Model with 4 variables could provide the best prediction result; variables selected are: Age, difftime, Gender, and Scholarship.

Cross validation to choose the best model

Final model

$$\log\left(\frac{p}{1-p}\right)$$

$$= -1.637 - 0.391 * I(Gender = M) + 0.401 * scholarship + 0.039 * difftime - 0.012 * Age$$

p is the probability that patient will not show up for the appointment.

No significant violation of model assumption was detected.

Cross validation to choose the best model

Model prediction

If the predicted probability is higher than 0.5, we predicted the response be ‘Yes’; otherwise we predict the response be ‘No’.

Then we compared the predicted response and real response and calculated the correct prediction rate.

For the original data, the correct prediction rate is 82%.

For data from some other cities, the correct prediction rates are also about 80%.

Discussion

- Cut off probability
- Correlation between variables
- Improve the tree: Boosting or Bagging

Acknowledgement

Prof. Loh

Prof. Fang

Questions?

Model assumption check