Medical Appointment Show Up

Context

A person makes a doctor appointment, receives all the instructions and no-show. Who to blame? Based on 110,000 medical appointment records, could we predict any patient will show up (***No Show***: *Yes / No*) for doctor appointment?

Refer ***here*** for source of data from Kaggle.

Exploratory Data Analysis

Initial Statistical Summary as follow for gender, scheduled & appointment date, age, scholarship, hypertension, diabetes, alcoholism, handicap, sms received and no show.

```
##     gender      schedule_date        appointment_date          age
##  Female:71840   Min.   :2015-11-10   Min.   :2016-04-29   Min.   : -1.00
##  Male  :38687   1st Qu.:2016-04-29   1st Qu.:2016-05-09   1st Qu.: 18.00
##                 Median :2016-05-10   Median :2016-05-18   Median : 37.00
##                 Mean   :2016-05-08   Mean   :2016-05-19   Mean   : 37.09
##                 3rd Qu.:2016-05-20   3rd Qu.:2016-05-31   3rd Qu.: 55.00
##                 Max.   :2016-06-08   Max.   :2016-06-08   Max.   :115.00
##  scholarship hypertension diabetes     alcoholism   handicap     sms_received
##  Yes:10861   Yes:21801    Yes:  7943   Yes:  3360   Yes:  2042   Yes:35482
##  No :99666   No :88726    No :102584   No :107167   No :108485   No :75045
##
##
##
##
##  no_show
##  Yes:22319
##  No :88208
##
##
##
##
```

Medical appointment records were scheduled (*scheduled_day*) between Nov 2015 & Jun 2016 and appointments will take place April - June 2016.

Notice age of patient range between -1 and 115. There are folks live up to 100+ year-old, however, -1 year-old certainly is not possible. Let's check this out.

```
## # A tibble: 1 x 6
##   patient_id        age appointment_id schedule_date appointment_date no_show
##   <chr>           <dbl> <chr>          <date>        <date>           <fct>
## 1 465943158731293    -1 5775010        2016-06-06    2016-06-06       No
```

The patient of -1 year-old probably just a pregnant lady who booked a doctor appointment. Will exclude this special case from analysis since I have no way to find out the actual age of the patient.

One thing got me curious is how long is the waiting period between scheduling a appointment until the appointment date.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -6.00    0.00    4.00   10.18   15.00  179.00
```

A simple summary above shows waiting time range from -6 days to 179 days. I will spend some time to figure out the negative waiting time.

```
## # A tibble: 23 x 6
##    patient_id appointment_id schedule_date appointment_date waiting_time no_show
##    <chr>      <chr>          <date>        <date>                  <dbl> <fct>
##  1 242522583~ 5657354        2016-05-04    2016-05-19                 15 Yes
##  2 242522583~ 5664962        2016-05-05    2016-05-04                 -1 Yes
##  3 378748196~ 5655637        2016-05-04    2016-05-03                 -1 Yes
##  4 378748196~ 5655638        2016-05-04    2016-05-10                  6 No
##  5 378748196~ 5655639        2016-05-04    2016-05-17                 13 No
##  6 378748196~ 5655642        2016-05-04    2016-05-24                 20 No
##  7 378748196~ 5655646        2016-05-04    2016-05-31                 27 No
##  8 783927266~ 5679978        2016-05-10    2016-05-09                 -1 Yes
##  9 783927266~ 5730318        2016-05-24    2016-05-24                  0 No
## 10 783927266~ 5752857        2016-05-31    2016-06-01                  1 No
## # i 13 more rows
```

Table above clearly show some appointments with negative waiting time. Interestingly, same patient could have records with positive & negative waiting time. Technically, system should have build in mechanism to make sure appointment date is on the same day or later than schedule date. Without much information related to such records, I will assume that appointments with negative waiting time are due to system issues. Therefore, I will not use such records for analysis / prediction.

For the appointments with 0 days waiting time, this is an indication that appointment and scheduling are on the same day.

Reproduce summary of appointment records.

```
##   waiting_time       gender           age         scholarship hypertension
##  Min.   :  0.00   Female:71836   Min.   :  0.00   Yes:10861   Yes:21801
##  1st Qu.:  0.00   Male  :38685   1st Qu.: 18.00   No :99660   No :88720
##  Median :  4.00                  Median : 37.00
##  Mean   : 10.18                  Mean   : 37.09
##  3rd Qu.: 15.00                  3rd Qu.: 55.00
##  Max.   :179.00                  Max.   :115.00
##  diabetes     alcoholism   handicap     sms_received no_show
##  Yes:  7943   Yes:  3360   Yes:  2040   Yes:35482    Yes:22314
##  No :102578   No :107161   No :108481   No :75039    No :88207
##
##
##
##
```

With the statistical summary, I may proceed to find out more about the most frequent waiting time for the appointments.

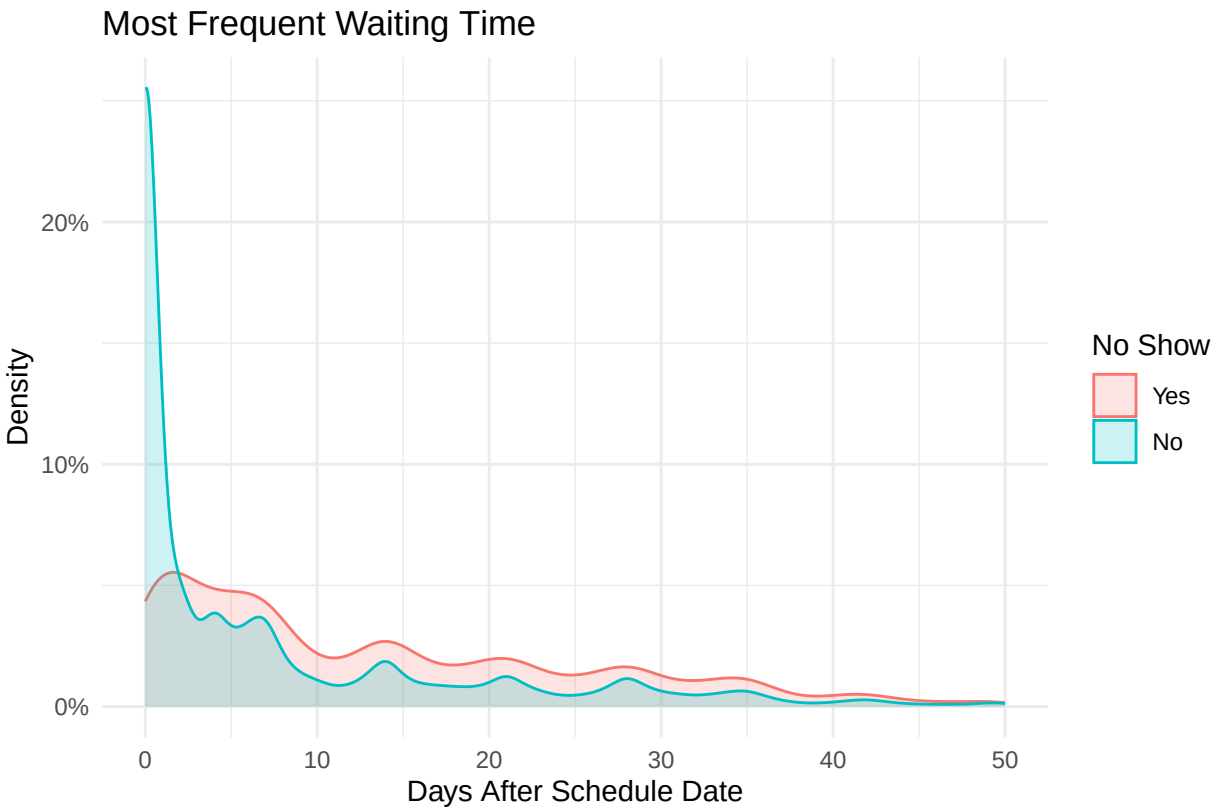

Figure 1

Observed from Figure 1, thought waiting time can range from 0 (same day) to 179 days (6 months), most appointments will either take place on the same day of scheduling the appointment or within 7 days after schedule date. This is especially true for patients who did show up (*No Show -> No*) for the appointments.

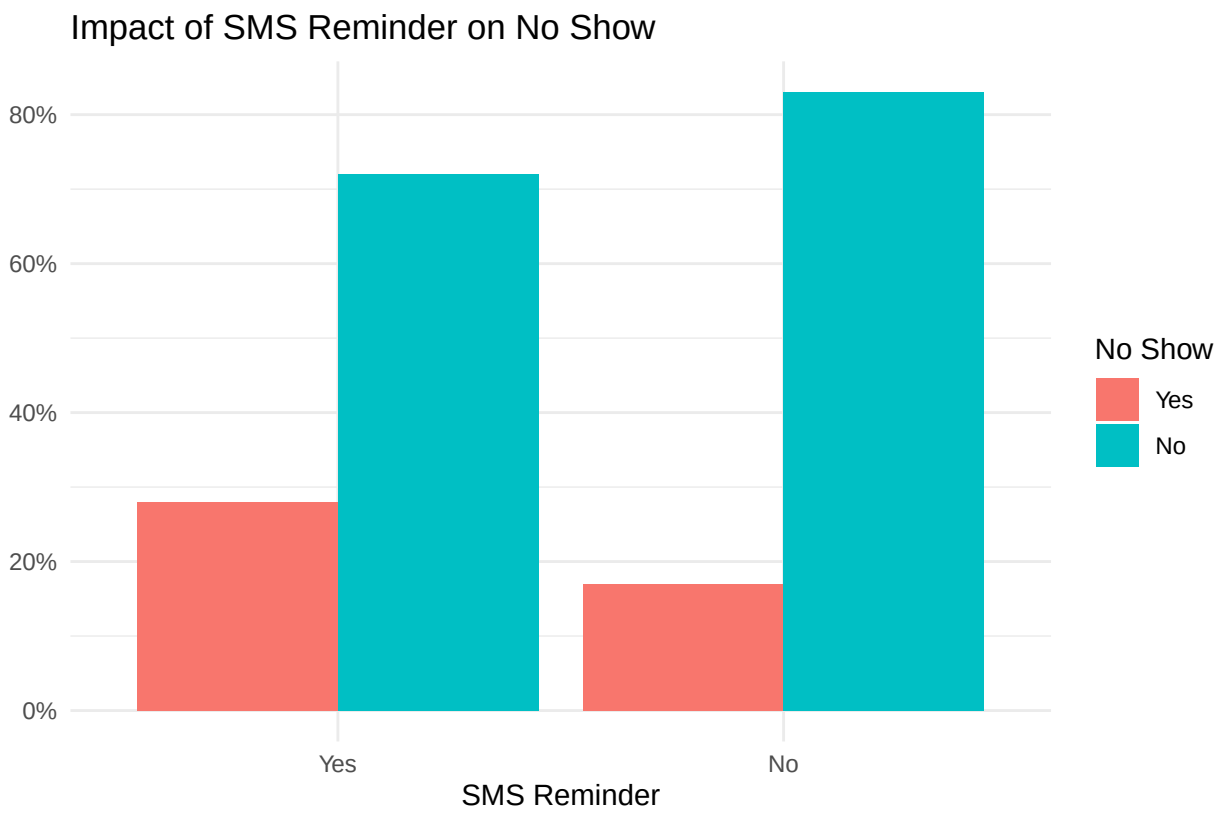

Figure 2

Logically, SMS reminder helps to remind patient for appointment. However, Figure 2 tell different story. Most patients ($> 60\%$) showed up for appointments (*No Show* == *No*), regardless they received SMS reminder.

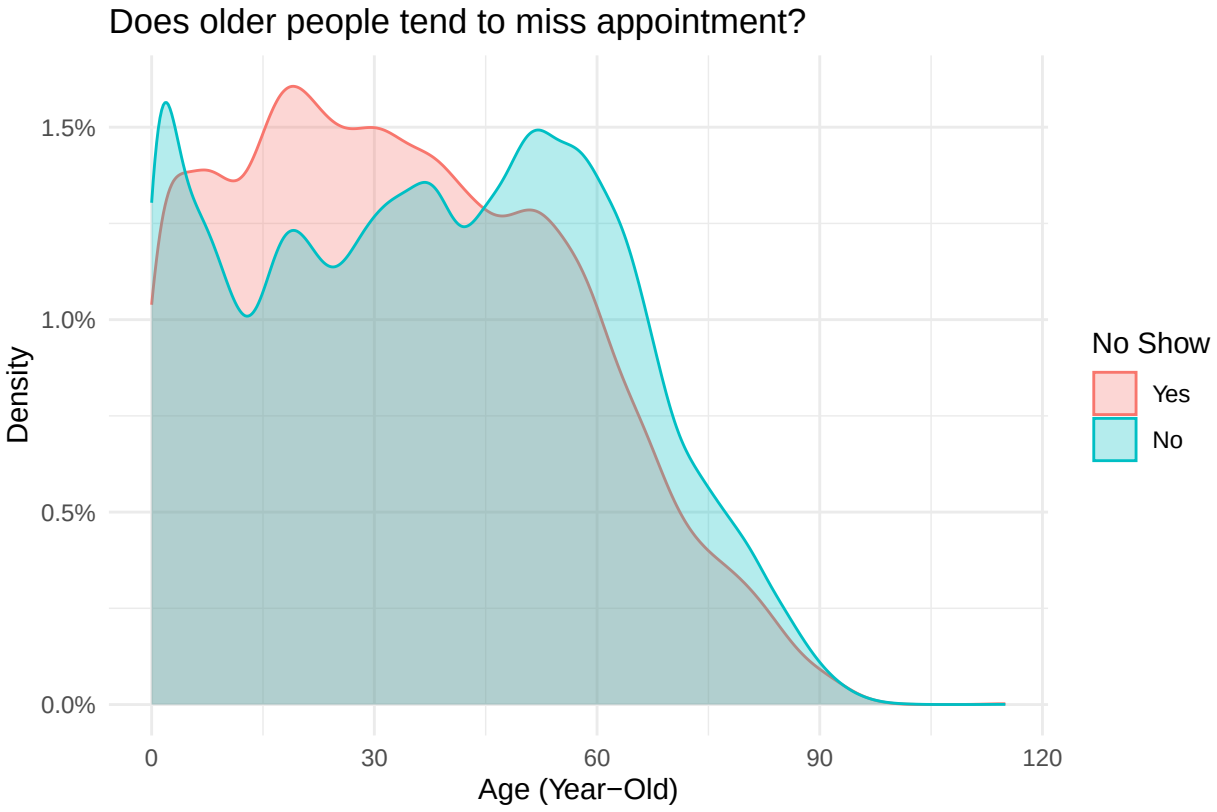

Figure 3

Based on Figure 3, people from different age group between 0 - 60 year-old primarily, are the majority of patients. However, there is no clear indication that patients who did not show up for appointments belong to certain age range.

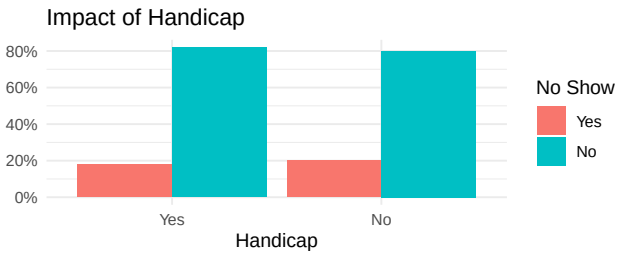

Figure 4

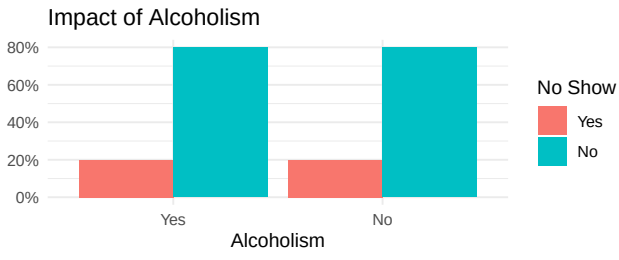

Figure 5

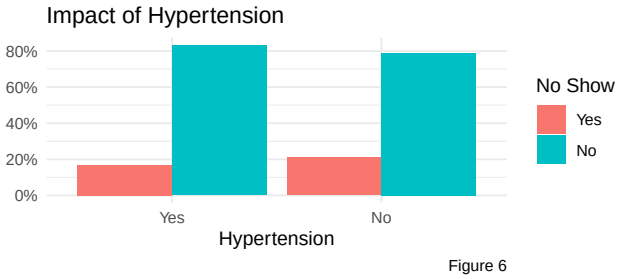

Figure 6

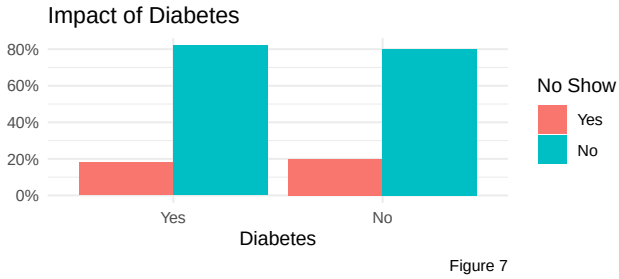

Figure 7

Figure 4, 5, 6 & 7 try to identify whether conditions such as handicap, alcoholism, hypertension and / or diabetes are in some ways preventing a patient from showing up for doctor appointment. If it does, I should expect to see opposite result between Yes & No given a condition. For example, by referring to Figure 4, the chart clearly shows no difference in outcome (No Show) given a patient is a handicap person; whether the patient is a handicap person, the percentage of No Show (No Show = Yes) is at ~ 20%.

Statistical Modelling with Logistic Regression

With all the basic analyses about waiting time, SMS reminder, age, impact of various health conditions, I will move on to predict whether a patient will show up for doctor appointment.

My hypothesis is that, older patient who has appointment months into the future, without sms reminder tend to forget about doctor appointment. Therefore, ***age***, ***waiting_time*** and ***sms_received*** should be the top predictors to tell whether a patient would show up (***no_show***) for doctor appointment.

However, I would first build a predictive model by including waiting_time, gender, age, scholarship, hypertension, diabetes, alcoholism, handicap and sms_received first then narrow down the selection of predictors later.

```
## # A tibble: 3 x 6
##   number operation type       trained skip  id
##    <int> <chr>     <chr>      <lgl>   <lgl> <chr>
## 1      1 step      downsample FALSE   TRUE  downsample_2fHgg
## 2      2 step      normalize  FALSE   FALSE normalize_y5nRw
## 3      3 step      dummy      FALSE   FALSE dummy_j2LVY
```

```
## # A tibble: 10 x 4
##    variable     type      role      source
##    <chr>        <list>    <chr>     <chr>
##  1 waiting_time <chr [2]> predictor original
##  2 gender       <chr [3]> predictor original
##  3 age          <chr [2]> predictor original
##  4 scholarship  <chr [3]> predictor original
##  5 hypertension <chr [3]> predictor original
##  6 diabetes     <chr [3]> predictor original
##  7 alcoholism   <chr [3]> predictor original
##  8 handicap     <chr [3]> predictor original
##  9 sms_received <chr [3]> predictor original
## 10 no_show      <chr [3]> outcome   original
```

Train logistic model and identify p-values for each relevant indicators.

```
## # A tibble: 6 x 5
##   term             estimate std.error statistic p.value
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>
## 1 waiting_time        0.639    0.0327    -13.7   0
## 2 age                 1.24     0.0308      7.01  0
## 3 sms_received_Yes    0.662    0.0586     -7.04  0
## 4 (Intercept)         1.21     0.0523      3.71  0.0002
## 5 scholarship_Yes     0.726    0.0883     -3.63  0.0003
## 6 alcoholism_Yes      1.53     0.134       3.18  0.0015
```

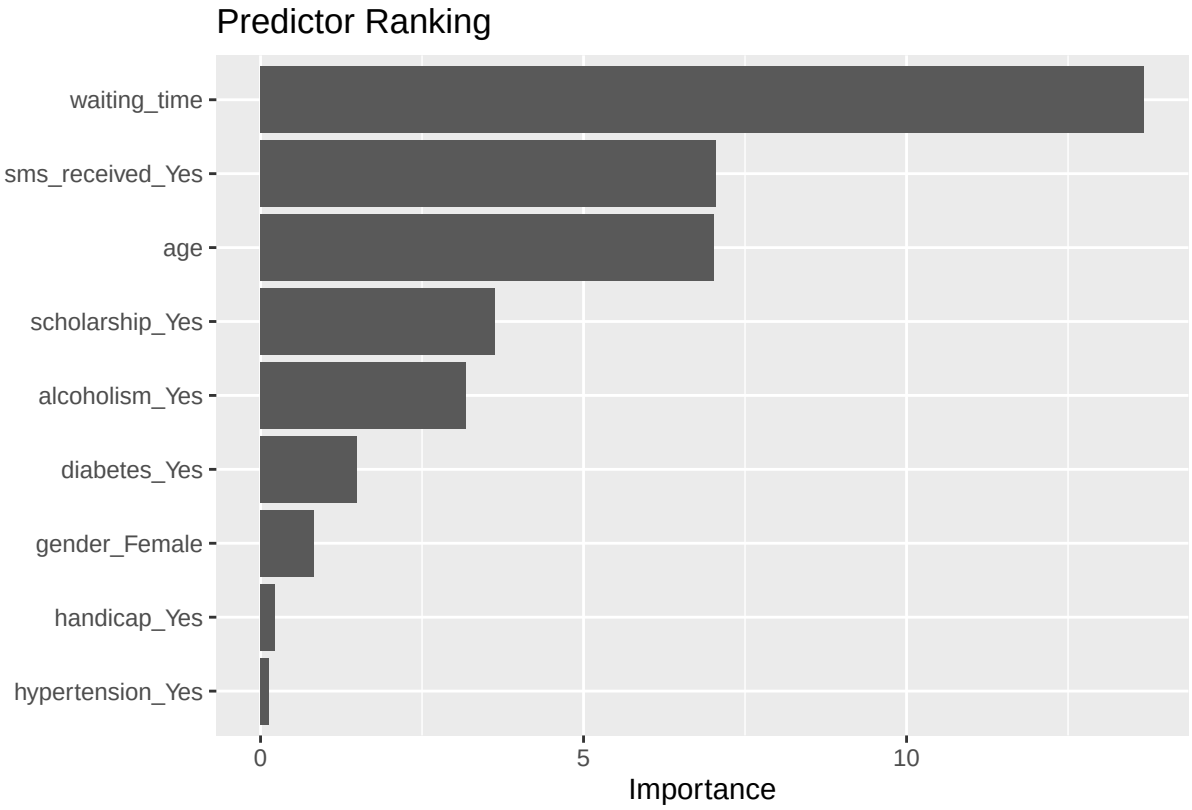

Figure 8

Trained model derived from 6000 patients appointment records show that waiting_time, age, sms_received, scholarship and alcoholism have the most impact on appointment no show. This is indicated by the p-value, which is below 0.05 (i.e: 95% confidence level). ***Predictor Ranking*** (Figure 8) is also align with the trained model statistic.

I will pick predictors with importance score above 5.0 to retrain the model: waiting time, age and sms_received.

```
## # A tibble: 4 x 4
##   variable     type      role      source
##   <chr>        <list>    <chr>     <chr>
## 1 waiting_time <chr [2]> predictor original
## 2 age          <chr [2]> predictor original
## 3 sms_received <chr [3]> predictor original
## 4 no_show      <chr [3]> outcome   original
```

Predict No Show with Model

Below is the statistical result after retrain logistic model with only waiting time, age and sms_received.

```
## # A tibble: 4 x 5
##   term             estimate std.error statistic p.value
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)         1.14     0.0342      3.74  0.0002
## 2 waiting_time        0.669    0.0317    -12.7   0
## 3 age                 1.21     0.0263      7.19  0
## 4 sms_received_Yes    0.691    0.0576     -6.41  0
```

p-values of the predictors remain below 0.05.

The coefficient values (refer to: ***estimate***) indicates the *impact* of each predictor on the outcome (*no_show*). The coefficient of ***age*** is *1.2084*. The positive value mean the *chance of a patient not showing up for an appointment increases as age increases*. The value itself indicates the magnitude of the impact; higher value mean greater impact. Same concept is applicable to waiting time and sms received, which mean patient has the tendency not to show up for appointment when waiting time is longer and he / she received SMS reminder.

Here is preview of new table includes actual outcome (no_show), waiting time, age, sms_received and the rest are predicted outcome & probabilities. ***.pred_class*** is the predicted outcome (Yes / No), ***.pred_Yes*** & ***.pred_No*** are the probability that appointment will be "Yes" & "No" respectively.

```
## # A tibble: 5 x 7
##   no_show waiting_time   age sms_received .pred_class .pred_Yes .pred_No
##   <fct>          <dbl> <dbl> <fct>        <fct>           <dbl>    <dbl>
## 1 No                 2    57 No           No              0.367    0.633
## 2 Yes                2    35 No           No              0.412    0.588
## 3 No                 2    62 No           No              0.357    0.643
## 4 No                 6    50 Yes          No              0.499    0.501
## 5 No                 2    64 No           No              0.353    0.647
```

All said, let's see how well my prediction is.

Below is confusion matrix that compares Actual (*Truth*) and *Prediction* for each outcome (Yes / No).

```
##           Truth
## Prediction  Yes   No
##        Yes 1776 3509
##        No  1158 7441
```

The model is applied to predict 13,884 appointments. Actual no show is 2,934 (Yes under Truth: 1,776 + 1,158) and total show up is 10,950 (No under Truth: 3,509 + 7,441). Predicted no show is 5,285 (Prediction with Yes: 1,776 + 3,509) and total show up is 8,599 (Prediction with No: 1,158 + 7,441).

Performance of my prediction is summarized as follow with metrics such as *Accuracy*, *Sensitivity* & *Specificity*.

```
## # A tibble: 6 x 3
##   .metric   .estimator .estimate
##   <chr>     <chr>          <dbl>
## 1 accuracy  binary         0.664
## 2 sens      binary         0.605
## 3 spec      binary         0.680
## 4 precision binary         0.336
## 5 f_meas    binary         0.432
## 6 mcc       binary         0.240
```

Metrics that worth attention:

1. *Accuracy = 0.6639*

I predicted 1,776 correctly for Yes (Yes for Truth & Prediction) and 7,441 correctly for No (No for Truth & Prediction) out of total 13,884 appointments. Overall prediction is about 66% accurate (1,776 + 7,441 correct / total 13,884).

2. *Sensitivity = 0.6053*

This metric only focus on accurate prediction for *Yes* outcome. Yes for Truth is 2,934 (Yes under Truth: 1,776 + 1,158) and my prediction recorded 1,776 as Yes. ~ 61% (1,776 / 2,934) of my prediction for Yes only is correct. The entire analysis is about predict whether a patient will NOT show up for appointment, hence my model accuracy is slightly higher than predicting Head or Tail of flipping a coin.

3. *Specificity = 0.6795*

This metric only focus on accurate prediction for *No* outcome. This is the opposite of Sensitivity metric. No for Truth is 10,950 (No under Truth: 3,509 + 7,441) and my prediction recorded 7,441 as No. ~ 68% (7,441 / 10,950) of my prediction for No only is correct. My model is doing better job when predicting a patient will show up (68%) for appointment instead of NOT showing up (61%).

We can visually inspect how well the predictive model perform.

From Figure 9 below, probability of NOT showing up trend higher as waiting time increases. Probability of NOT showing up is higher for the appointments where the patients received SMS reminder given the same waiting time.

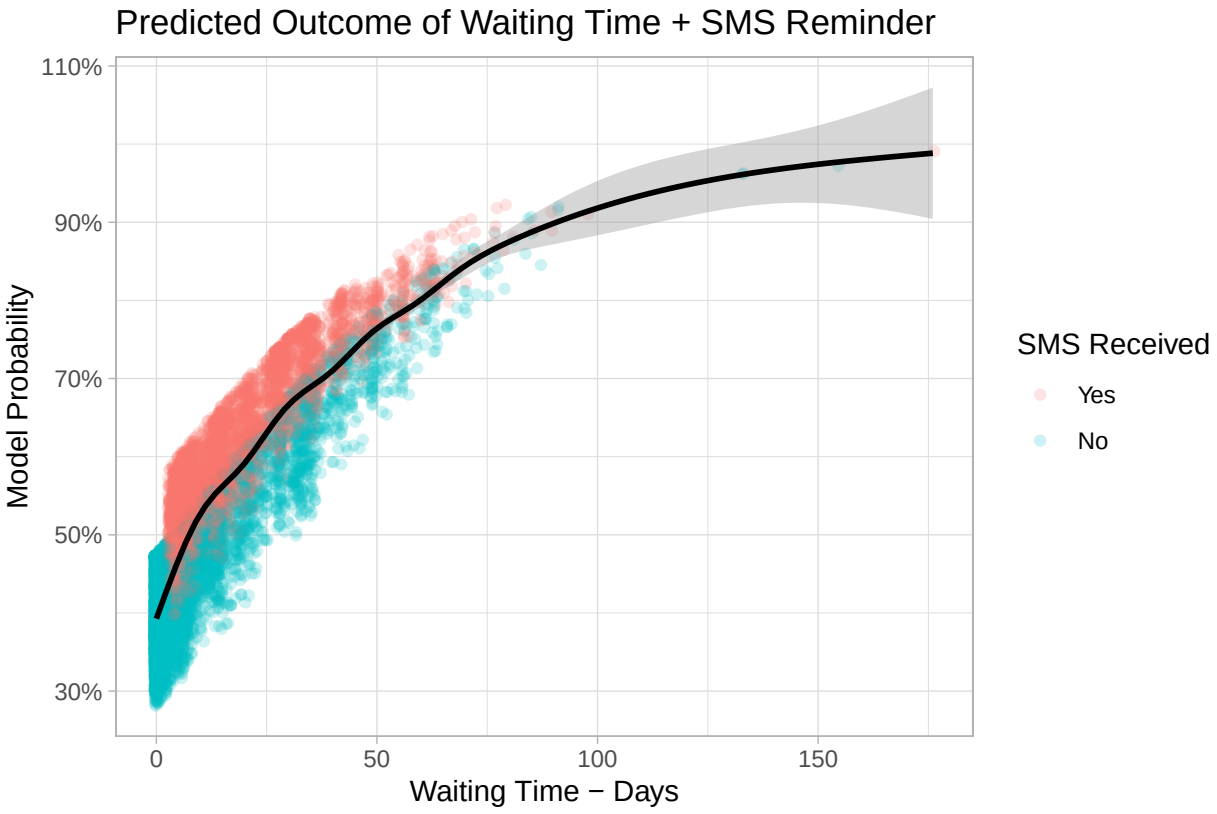

Figure 9

However, Figure 10 below shows degradation of model performance when age is included as part of the predictive model on top of waiting time & sms received (refer to left chart of Figure 10). The direction of trend line for No Show = Yes is not obvious.


Predicted Outcome as Age & Waiting Time Increases with SMS Reminder
Yes
No
Waiting Time – Days
150
100
50
0
0
25
50
75
100
Age [Year Old]
No Show
Yes
No

Figure 10