Missed appointment analysis

Ravshan S.K.

December 6, 2018

Introduction

We were given a dataset of online medical appointments in the city of Vitória, Espírito Santo in Brazil. It turned out that in a period of three months between May and August, 2016, patients did not show up in 25% of appointments. To investigate the possible reasons behind this, we will analyze the data and make inferences.

Data Exploration

Outliers

First of all, we explore the dataset. We immediately notice that some age values are negative, and very old patients don't exhibit variation having too few observations:

```
table(df$age)
```

```
##
##    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13
##     1  3539  2273  1618  1513  1299  1489  1521  1427  1424  1372  1274  1195  1092  1103
##    14    15    16    17    18    19    20    21    22    23    24    25    26    27    28
##  1118  1211  1402  1509  1487  1545  1437  1452  1376  1349  1242  1332  1283  1377  1448
##    29    30    31    32    33    34    35    36    37    38    39    40    41    42    43
##  1403  1521  1439  1505  1524  1526  1378  1580  1533  1629  1536  1402  1346  1272  1344
##    44    45    46    47    48    49    50    51    52    53    54    55    56    57    58
##  1487  1453  1460  1394  1399  1652  1613  1567  1746  1651  1530  1425  1635  1603  1469
##    59    60    61    62    63    64    65    66    67    68    69    70    71    72    73
##  1624  1411  1343  1312  1374  1331  1101  1187   973  1012   832   724   695   615   725
##    74    75    76    77    78    79    80    81    82    83    84    85    86    87    88
##   602   544   571   527   541   390   511   434   392   280   311   275   260   184   126
##    89    90    91    92    93    94    95    96    97    98    99   100   102   115
##   173   109    66    86    53    33    24    17    11     6     1     4     2     5
```

Moreover, some appointments have been done to the dates before it was scheduled, probably, due to some system error.

```
table(as.numeric(df$dayap - df$daysc)<0)
```

```
##
##  FALSE   TRUE
## 110522      5
```

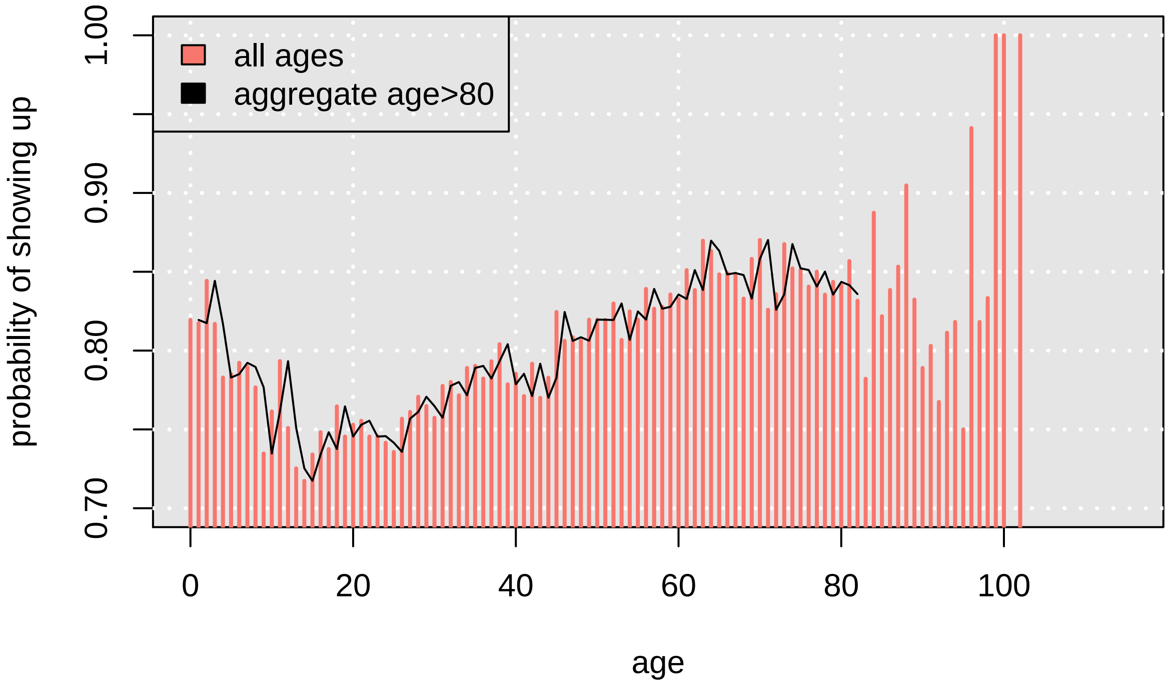
We remove the obvious outliers from our data, combine all data points for households older than 80 (to balance subset size), and continue with our analysis.

```
df <- df[ (df$age >= 0) & (df$dayap >= df$daysc),]
```

Demographic factors

Age factor

Observing the no-show dynamics throughout lifetime, gives important insights to our analysis:



From common experience, we know that appointments for patients younger than 18, are actually done by their parents. We can see how, as kids grow older, the parents tend to miss more appointments (because new parents tend to be concerned more with infant's health, and as kids grow, parents tend to neglect their health slightly more).

As kids grow into young adults, they start steadily taking their health seriously and miss less appointments.

Also, for patients older than 80, we don't have many data points at particular ages, so we aggregated them into one group.

So, we use two possible age variables: a continuous one with two dummies for 18 and 80 year olds, and a discrete variable correpsonding to 11 decades of patients. #### Gender factor Another demographic factor is gender. The probability distribution shows that gender does not play a significant role in no-show rate:

```
##             show    noshow
## female 0.7968846 0.2031154
## male   0.8003619 0.1996381
```

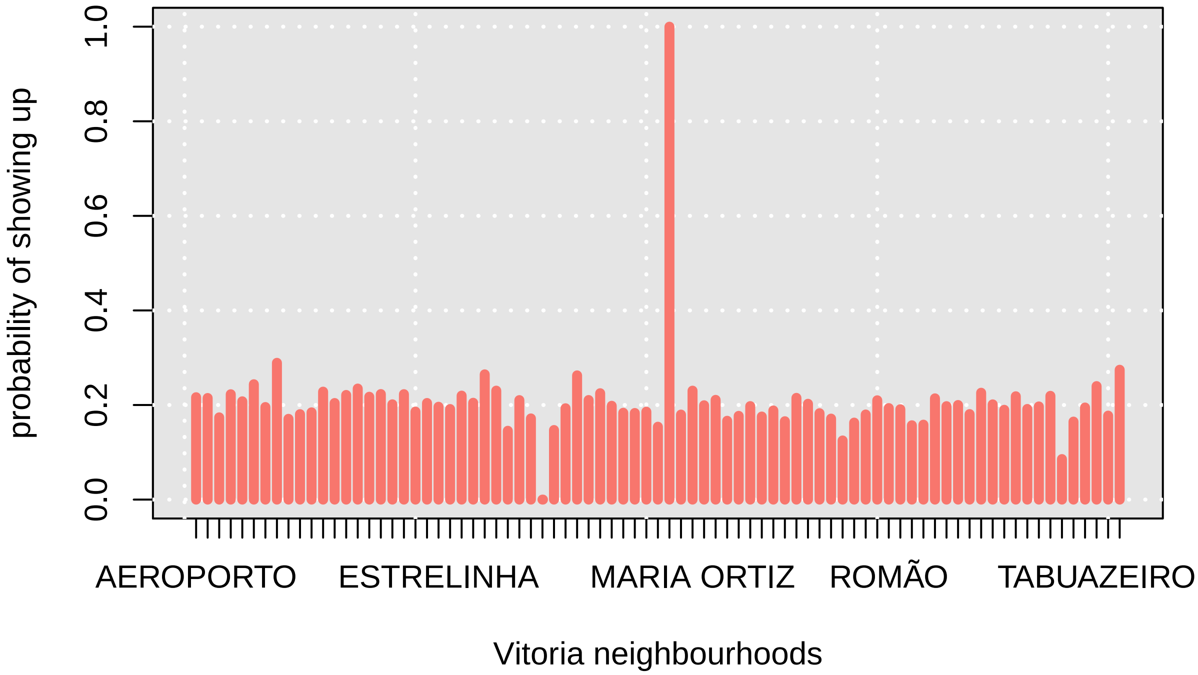
We tried to look at adults only (age>=18), because usually mothers (female=1) go to doctors with their children, irrespective of the kid's registered gender, but this procedure returned no significant difference either:

```
##             show    noshow
## female 0.8010114 0.1989886
## male   0.8099980 0.1900020
```

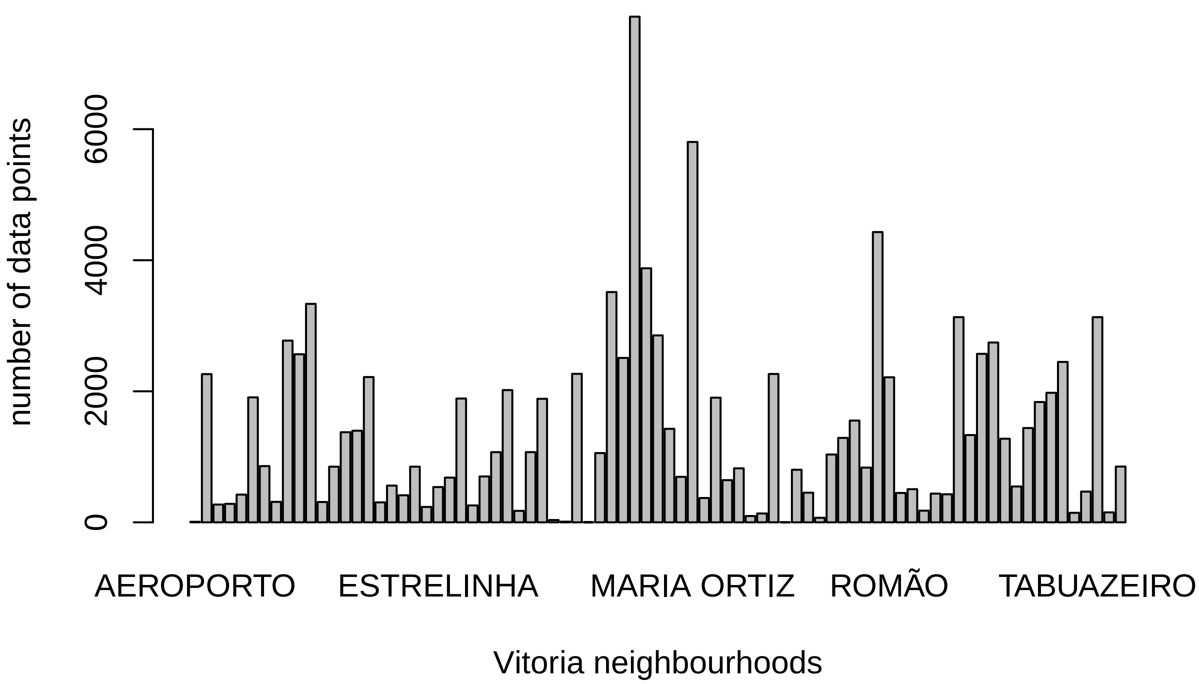
We ignore gender in further discussion.

Geographic factor

Looking at no-show rates by neighbourhood shows relative balance, with only a few outliers, which do not round to 20%.



We see that most (not all) of the outliers come from neighbourhoods with little data:



Therefore, we ignore neighbourhoods with less than 40 data points to avoid wrong statistics (e.g. in Parque Industrial there is only one registered appointment and it shows 100% show-up rate, which is incomparable with neighbourhoods with thousands of observations.) So, we ignore neigbourhoods *Aeroporto, Ilha do Boi, Ilha do Frade, Ilhas Oceanicas de Trinade, and Parque Industrial.*

Now, we still have outliers (we considered 3% to be a significant deviation from the average, 20%) in percentage of no-shows, which have sufficient observations not to ignore them. They are: *Solon Borges, Santos Dumont, Santa Clara, Santa Cecilia, Itarare, De Lourdes, Do Cabral, Do Quadro, Horto, Jardim Da Penha, Jesus de Nazareth, Mario Cypreste, and Santa Martha.*

We will add dummy variables for these 13 neighbourhoods as our geographic predictors. Other neighborhoods will be assumed to contribute no new information to the expected value.

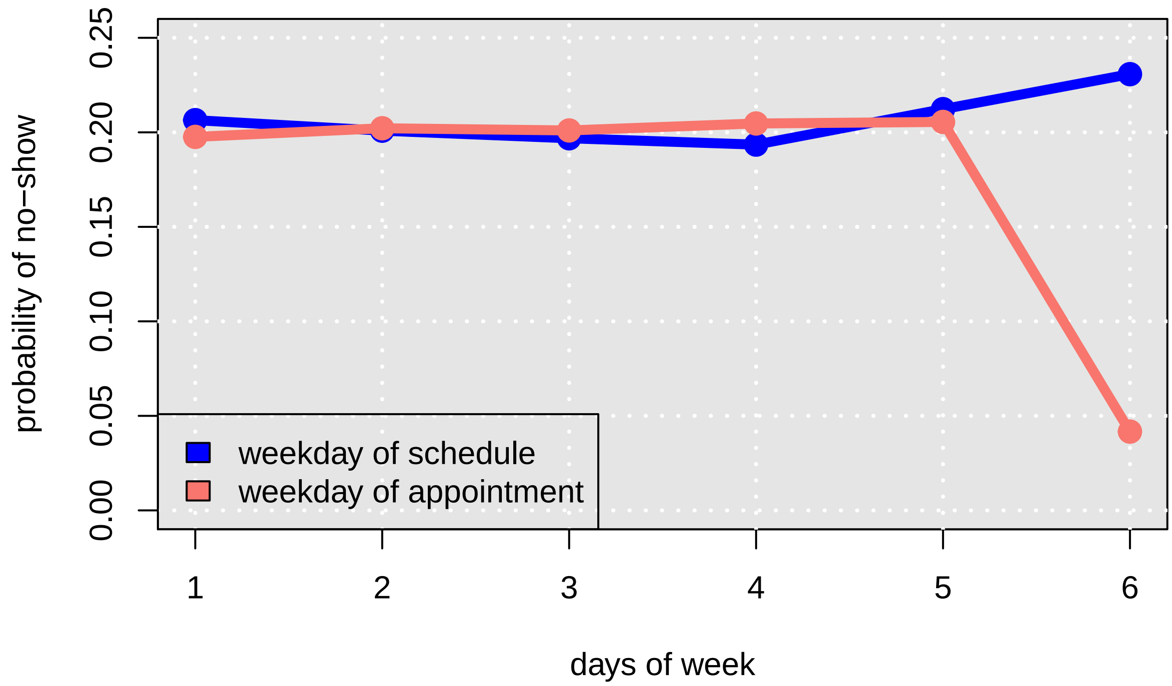
```
df$solbor <- ifelse(df$rayon=="SOLON BORGES",1,0)
df$sandum <- ifelse(df$rayon=="SANTOS DUMONT",1,0)
df$sancla <- ifelse(df$rayon=="SANTA CLARA",1,0)
df$sancec <- ifelse(df$rayon=="SANTA CECÍLIA",1,0)
df$itarar <- ifelse(df$rayon=="ITARARÉ",1,0)
df$lourde <- ifelse(df$rayon=="DE LOURDES",1,0)
df$cabral <- ifelse(df$rayon=="DO CABRAL",1,0)
df$quadro <- ifelse(df$rayon=="DO QUADRO",1,0)
df$horto  <- ifelse(df$rayon=="HORTO",1,0)
df$penha  <- ifelse(df$rayon=="JARDIM DA PENHA",1,0)
df$jesus  <- ifelse(df$rayon=="JESUS DE NAZARETH",1,0)
df$cypres <- ifelse(df$rayon=="MÁRIO CYPRESTE",1,0)
df$sanmar <- ifelse(df$rayon=="SANTA MARTHA",1,0)
```

Temporal factor

Temporal data brings the crucial information about the appointment no-shows, starting from the weekday of the appointment, and ending with the wait time. Since we don't have at least a year-long data, we cannot speak of seasonality patterns, and will have to get by with what we have.

Weekdays

The day of the week is the first thing that comes to mind - during the weekdays, patients might have emergencies at school or at work, and this could cause them to miss the appointment. However, the analysis shows no significant difference through week, except for Saturdays:



But further analysis shows that this happens because of the lack of enough data for Saturday:

```
table(df$wdayap,df$noshow)
```

```
##
##         0     1
##   1 18024  4689
##   2 20488  5150
##   3 20774  5092
##   4 13909  3337
##   5 14982  4037
##   6    30     9
```

```
table(df$wdaysc,df$noshow)
```

```
##
##         0     1
##   1 18523  4561
##   2 20877  5290
##   3 19383  4876
##   4 14373  3699
##   5 15028  3887
##   6    23     1
```

So, we ignore the weekdays and assume that they don't affect the no-show rate.

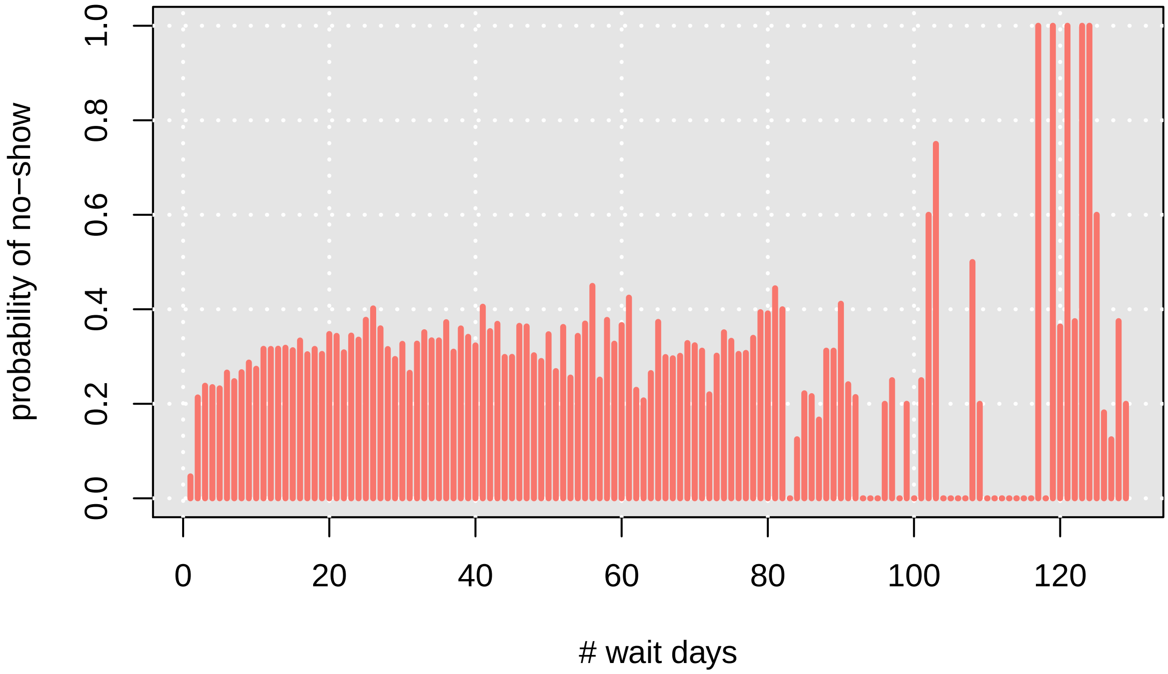
Wait time

Another obvious variable is a wait time from scheduling the appointment to the appointment itself. It is plausible to assume that in longer wait times patients can get cured, book another earlier appointment, or die before the appointment.

We define the "days between" variable as the difference between "day of appointment" and "day of schedule":

```r
df$daybw <- as.numeric(df$dayap - df$daysc)
```

The plot below shows that when appointments are scheduled in that same day (wait time = 0), the patient almost never misses it (just 4% no-show rate). There is sufficient data (34% of all observations) to support this claim.



In the next days, the no-show rate is very volatile. To avoid weekly seasonality, we aggregate the data by weeks. Also, for two reasons: *(i)* since the longer wait times have small data points and *(ii)* since (assuming time discounting — a weak economic assumption) people perceive recent past clearer than distant past, we aggregated first month as 4 weeks, and aggregated the next data points by month:

```
##               show    no show
## 0 days   0.9535294 0.04647062
## 1 week   0.7585211 0.24147895
## 2 weeks  0.6953015 0.30469854
## 3 weeks  0.6775975 0.32240252
## 4 weeks  0.6633353 0.33666468
## 2 months 0.6662968 0.33370322
## 3 months 0.7034588 0.29654120
## 4 months 0.7363014 0.26369863
```

Note that *4 months* wait is an outlier due to small dataset. Otherwise, all probabilities after 1 week fall into ±3% interval. Taking this and *(ii)* into consideration, suggests an even wilder (yet still plausible) aggregation: *0 days, 1 week, and >1 week*:

```
##              show    no show
## 0 days  0.9535294 0.04647062
## 1 week  0.7585211 0.24147895
## >1 week 0.6794388 0.32056117
```

Thus, we will use three-valued categorical variables to denote wait times.

Hour of the day

Lifestyle of people potentially reflects their degree of responsibility — *"night owls"* tend to sleep during days and maybe miss deadlines, while *"early birds"* may take their appointments more seriously. We take a look at the data of time o'clock when the appointment was scheduled. The online appointment system opens at 6AM and closes at 10PM. We divide these 16-hour days into 4 groups of 4, and find that *"early bird effect"* actually exists, and people who scheduled appointments between 6AM and 10AM are significantly less likely to miss their appointments, while any other time slot does not change the no-show probability significantly:

```
##                 show   no show
## 6AM-10AM: 0.8255142 0.1744858
## 10AM-2PM: 0.7833574 0.2166426
## 2PM-6PM:  0.7668771 0.2331229
## 6PM-10PM: 0.7773175 0.2226825
```

Thus, we use a binary dummy variable — *6AM-10AM* or *10AM-10PM* — to incorporate time.

Medical factor

Appointment history

To incorporate the idiosyncracies of patients, we use the history of their previous appointments, and whether they missed them before. The table below shows the repeat patients' allocation.

```
##
##     0     1     2     3     4     5     6     7     8     9    10    11
## 62295 24378 10484  4984  2616  1498   945   639   437   333   248   185
##    12    13    14    15    16    17    18    19    20    21    22    23
##   149   114    92    77    67    57    49    43    35    32    31    29
##    24    25    26    27    28    29    30    31    32    33    34    35
##    28    28    28    28    28    27    25    25    25    24    22    21
##    36    37    38    39    40    41    42    43    44    45    46    47
##    21    20    18    18    17    17    15    15    15    15    13    13
##    48    49    50    51    52    53    54    55    56    57    58    59
##    13    13    12    11    11    11    10     9     9     8     8     8
##    60    61    62    63    64    65    66    67    68    69    70    71
##     8     8     4     4     4     3     3     3     3     3     2     2
##    72    73    74    75    76    77    78    79    80    81    82    83
##     2     2     2     2     2     2     2     2     2     2     2     2
##    84    85    86    87
##     1     1     1     1
```

For every patient, we found the modes (most frequent observations) of previous no-show stats. For patients that appear in the dataset only once, this value will be 0 (performing sensitivity checks we learned that leaving out one-time patients entirely, returns no significant difference (<0.5%) but complicates the dataset, so we chose to set the average previous no-show rate for first-time patients at 0). We observe that such modes greatly contribute to predicting the next no-show:

```
##              show    noshow
## mode0: 0.8116890 0.1883110
## mode1: 0.6652691 0.3347309
```

Patient condition

Patient's medical history can influence the no-show behavior. The tables below show differences in percentage of no-shows for patients with alcoholism, diabetes, hipertension, medical financial assitance (so-called "scholarship"), and handicaps:

```
##                      show   no show
## no scholarship 0.8019667 0.1980333
## scholarship    0.7626370 0.2373630
```

```
##                       show   no show
## no hipertension 0.7910054 0.2089946
## hipertension    0.8269804 0.1730196
```

```
##                   show   no show
## no diabetes 0.7964086 0.2035914
## diabetes    0.8199673 0.1800327
```

```
##                  show   no show
## no alcohol 0.7980889 0.2019111
## alcohol    0.7985119 0.2014881
```

```
##                  show   no show
## handicap0: 0.7976672 0.2023328
## handicap1: 0.8215686 0.1784314
## handicap2: 0.7978142 0.2021858
## handicap3: 0.7692308 0.2307692
## handicap4: 0.6666667 0.3333333
```

Strangely, alcoholism is the only condition which turned out to not have an effect on no-show probability. The probable reason is that alcoholism is not immediately lethal, and people tend to treat is less seriously than any other *"more serious"* illness like diabetes. So, we include all of the above conditions, except alcoholism, in our classification.

SMS

One can justifiably argue that patients may simply forget their appointment date and time. Hospitals tried to send SMS-reminders to their patients, but does this practice worth the cost? A simple frequency table shows that yes, patients, who received SMS-reminders, were less likely to miss the appointment:

```
##               show    no show
## no sms 0.56558482 0.11337212
## sms    0.23251690 0.08852616
```

Overall no-show stata

Finally, we decided to include the general historical average of noshows till the moment by all patients. However, this complicated the random sampling, and, most importantly, didn't affect the final result much (because it had little variation over time). So, we did not include this variable.

Classifier

We used *logit* (logistic regression) and *decision trees* to classify the data.

Logit

```
logit <- glm(formula =
               noshow ~ age + age18 + age80 +
               burs + diabet + handcap + alcohol +
               solbor + sandum + itarar + lourde + cabral + quadro + penha + jesus + sanmar +
```

```r
               sms + morning + waittime + mode_previous,
             family = binomial(link="logit"),
             data = dftrain
             )
summary(logit)
```

```
##
## Call:
## glm(formula = noshow ~ age + age18 + age80 + burs + diabet +
##     handcap + alcohol + solbor + sandum + itarar + lourde + cabral +
##     quadro + penha + jesus + sanmar + sms + morning + waittime +
##     mode_previous, family = binomial(link = "logit"), data = dftrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5490  -0.7291  -0.4477  -0.3346   2.6454
##
## Coefficients:
##                 Estimate Std. Error z value             Pr(>|z|)
## (Intercept)   -2.3027581  0.0267987 -85.928 < 0.0000000000000002 ***
## age           -0.0125877  0.0006706 -18.772 < 0.0000000000000002 ***
## age18          0.2659647  0.0312217   8.519 < 0.0000000000000002 ***
## age80          0.3157662  0.0606324   5.208  0.00000019101044335 ***
## burs           0.1972190  0.0286774   6.877  0.00000000000610610 ***
## diabet         0.1544029  0.0372166   4.149  0.00003342670672459 ***
## handcap        0.0915749  0.0563530   1.625             0.104158
## alcohol        0.3330153  0.0522980   6.368  0.00000000019194428 ***
## solbor        -0.4963353  0.1532160  -3.239             0.001198 **
## sandum         0.2988702  0.0764425   3.910  0.00009239634926811 ***
## itarar         0.2022540  0.0464296   4.356  0.00001323766938002 ***
## lourde        -0.5588738  0.1891486  -2.955             0.003130 **
## cabral        -0.3399142  0.1339569  -2.537             0.011165 *
## quadro        -0.2988064  0.1078614  -2.770             0.005601 **
## penha         -0.3093562  0.0516199  -5.993  0.00000000206043028 ***
## jesus          0.1934662  0.0537897   3.597             0.000322 ***
## sanmar        -0.1660085  0.0588319  -2.822             0.004776 **
## sms           -0.1569562  0.0200545  -7.826  0.00000000000000502 ***
## morning       -0.1422520  0.0177617  -8.009  0.00000000000000116 ***
## waittime       0.9883130  0.0130948  75.474 < 0.0000000000000002 ***
## mode_previous  0.7399796  0.0268507  27.559 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88976  on 88416  degrees of freedom
## Residual deviance: 79780  on 88396  degrees of freedom
## AIC: 79822
##
## Number of Fisher Scoring iterations: 5
```

We found that the wait time, previous no-show history, and geographical location were the most important predictors. Also, alcohol turned out to be statistically significant, while hipertension had to be removed from the final model. Strangely, eligibility to financial assistance increased the likelihood of no-show (but this was discussed in the previous section).

To check the accuracy of the model, we predict the no-show for the test data and use measures called *accuracy* and *AUR*:

```
predicted_noshow <- ifelse(predict(logit, dftest,type = "response")>=0.5,1,0)
actual_noshow <- dftest$noshow
mytab <- table(abs(predicted_noshow - actual_noshow))

accuracy <- mytab[1]/(mytab[1] + mytab[2])
accuracy
```

```
##         0
## 0.8000362
```

```
roc(actual_noshow, predicted_noshow)
```

```
##
## Call:
## roc.default(response = actual_noshow, predictor = predicted_noshow)
##
## Data: predicted_noshow in 17650 controls (actual_noshow 0) < 4454 cases (actual_noshow 1).
## Area under the curve: 0.5231
```

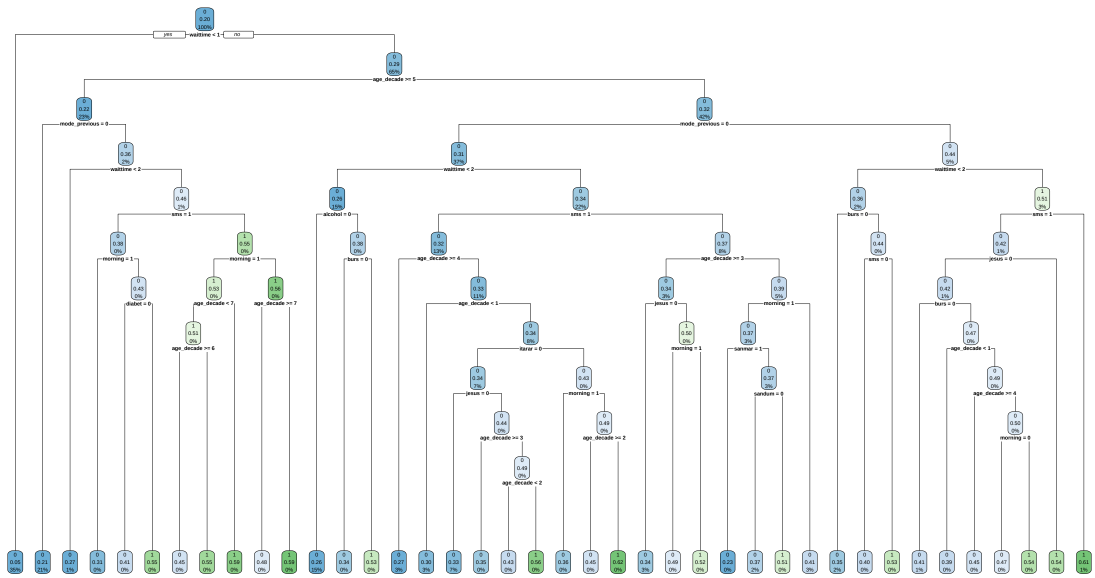
So, we have 81% accuracy, but this is not a very desirable result because of data imbalance — if we just set all "noshows" to zero, we would still get 80% accuracy. The measure *AUR* confirms this by giving the result 52.9% — which is slightly better than saying *"fifty-fify"*.

So, we decide to take a look at decision trees:

Decision trees

The decision tree returns 79% accuracy and 50.6% *AUR* on average. Different tree specifications give different resulting trees, but the average accuracy doesn't change. Here is just one of the trees. Notice that we used age as decades here, and not as a continuous variable:

```
dtree <- rpart(formula=noshow ~ age_decade +
                 burs + diabet + handcap + alcohol + hiper +
                 solbor + sandum + itarar + lourde + cabral + quadro + penha + jesus + sanmar +
                 sms + morning + waittime + mode_previous,
               data = df,
               method = "class",
               control = rpart.control(xval = 5, cp=0.0000001, minsplit=100)
               )
rpart.plot(dtree)
```



Conclusion

We have analyzed the data and tried to do a classification analysis. The results are not perfectly accurate, but they are not worse than a naive “noshow=0” prediction.

As an economist, I believe that having a higher accuracy would be impossible without additional data, like weather that day (was it rainy or not), the traffic situation, the diagnosis, the local news etc.

And here, we conclude.