Data Science Project:

Time is efficiency: a study of no-show in healthcare

I. Scope and plan of study

In 2017, the NHS put forth a ten-point plan to reduce efficiency losses within its healthcare system[1]. Indeed, efficiency maximization is a subject that concerns all domains. In healthcare, the increasing demand side pressure by an aging population means that tiny inefficiencies translate in large costs - Jacobs et al(2006). One such case is the sub-optimal turn-up rate at medical appointments. The heart of the problem lies in minimizing doctor idle time and patient waiting time while considering no-shows and unpunctuality. The issue is not a recent one; Bean et Talage (1995) through empirical proof showed that factors like the relation the patient has with the doctor (i.e. recurring or new) and other factors such as age or gender had an impact on appointment cancellation rates. This problematic has been studied in many ways such as the effect of reminders in no-show minimization – McLean et al. (2016) - or finding the perfect overlapping schedule time –Anderson et al (2014). As pointed out by Javid et al (2017), optimisation of scheduling has gotten more attention in recent years and has lead to state of the art models such as the one developed by Deceuninck et al. (2017) which classifies patients by characteristics and optimizes patient sequencing. In this context, this empirical study aims to add to this body of literature by studying 'No-show' rates in hospitals in Vitoria – Brazil - in 2016. The dataset was collected from Kaggle.com, which obtained it from the municipality of Vitoria. The aim of the study is two-fold: create a robust model to predict the 'No-show' rate in Vitoria (predictive analysis) and reflect with the literature on the main causes of 'No-show' (inference discussion). The study method used is a trained and tested logistic regression and the analysis of its coefficients that leads to inference. The questions considered are the influence of age, gender, appointment time and patient no-show history (along with certain controls) on the probability of not turning up to a medical appointment. Each of these features could be studied in depth. However, to narrow the scope of this paper, we shall evaluate more closely how the appointment time gives us insight on the probability of a patient not showing up to an appointment.

As the problem at hand is a binary classification, the strategy to construct and evaluate the model is the following:

1. Tidy and transform the data appropriately (here particular attention is paid on the transformation of categorical variables as they are crucial to the interpretation of the model)

2. Divide the data between a training set, a validation set and a test set. From here onwards, we put aside the test set and construct the model on the training and validation sets.

[1] https://www.england.nhs.uk/five-year-forward-view/next-steps-on-the-nhs-five-year-forward-view/funding-and-efficiency/

3. Discuss the coefficients and their effects, if any, by performing hypothesis testing and select the best model.

4. Fit the selected model on the test set and verify how well it predicts the test set.

II. Analytical process

A. Data transformations

In this section, we detail the pre-processing steps taken before fitting the model. The main tools used were the Python packages pandas, numpy, scikit-learn, and seaborn. This choice was made as the analysis did not require any tools domain specific and all steps including pre-processing and visualisation could be accomplished through this mean.

The first data preparation step was the evaluation and reflection on missing values. Fortunately, available dataset was cleaned and only included one observation which could be qualified as a measurement error. This observation indicated a negative Age. Since this was the only visible error in a large dataset, it was deleted.

An important step in this study was the data transformations necessary to the analysis. Indeed, as most of the variables were categorical in nature, the transformations in this step were paramount in the interpretation of any results. The first objective was to make sure all categorical variables were numerical and interpretable. The following transformations were all done through pandas:

- The feature Gender was transformed from a string to a simple dummy variable such that Gender = 1 if the individual was male and 0 otherwise.
- The feature Handicap was originally equal to the number of different handicaps an individual had (i.e. the max value was 4 for an individual with 4 different reported handicaps). However, it was decided that a binary variable where Handicap = 1 if the person was handicapped and 0 otherwise made more sense due to the overwhelming number of people with a single handicap compared to those with more than one. Also, the unavailability of the list of different handicap situations created ambiguity in the variable that would be better described in dummy variable.
- The feature Neighbourhood was a string that indicated where the patient was from. As there are many neighbourhoods in Vitoria (82 reported in the data), it would not have been feasible or interesting to create a category per neighbourhood. Also, due to scarcity of additional information of the neighbourhoods (such as exact geographic location or population) it was decided to create categories that would accurately reflect the distribution of patients per neighbourhood (displayed in figure 1).

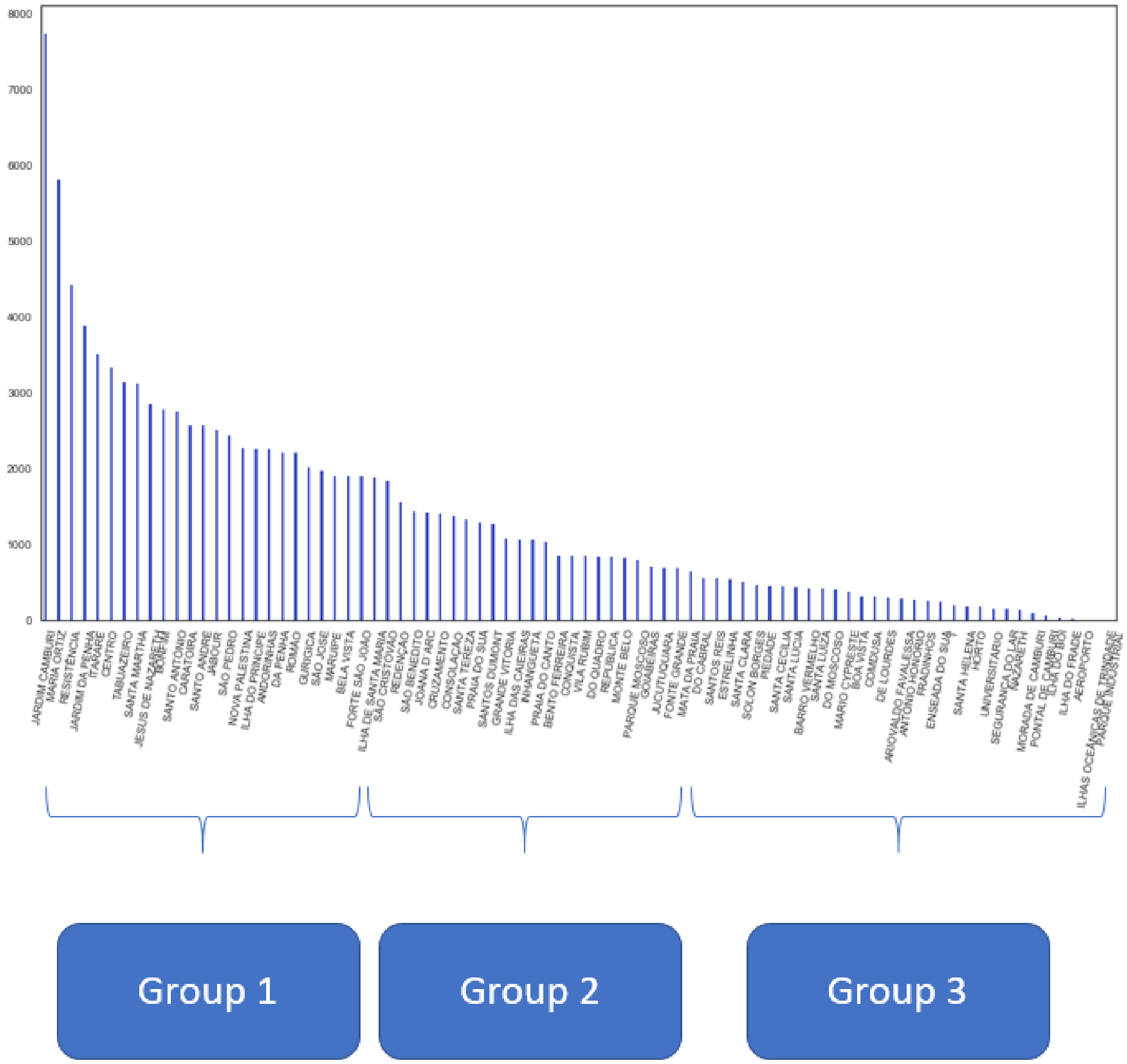


Figure 1 - Count of patients by neighbourhood

Consequently, three groups were created where:

$$Neighbourhood_i = \begin{cases} 2 \; if \; patient \; i \; in \; group \; 1 \\ 1 \; if \; patient \; i \; in \; group \; 2 \\ \quad 0 \; otherwise \end{cases}$$

Where group 1 contains the first 26 neighbourhoods in terms of patient count, group 2 the following 26 and group 3 the remaining 30.

- The features ScheduledDay and AppointmentDay both respectively described, through strings, the moment the doctor appointment was taken and the moment of the actual appointment in the format ‘YYYY-MM-DD [time]’. As one of the features of interest is the amount of time between the appointment day and the moment it was scheduled, the original features were transformed into timestamps and then a delta time variable called Time was defined as:

$Time_i = AppointmentDay_i - SchedulededDay_i$ for patient i

Interestingly, there was a measurement error that came to light at this stage: around 35% of the Time feature had negative values. As this proportion was too large to drop, a mean replacement was applied. This explains the distribution of Time displayed in figure 2. To be integrated in the regression only the day component of the delta time type variable was kept as an integer. Another feature of interest is the day on which patients scheduled the appointment for. To extract this information, the pandas function dayofweek was used to create a categorical variable Days where Days $= 0$ if the appointment was on Monday up till 6 for Sunday.

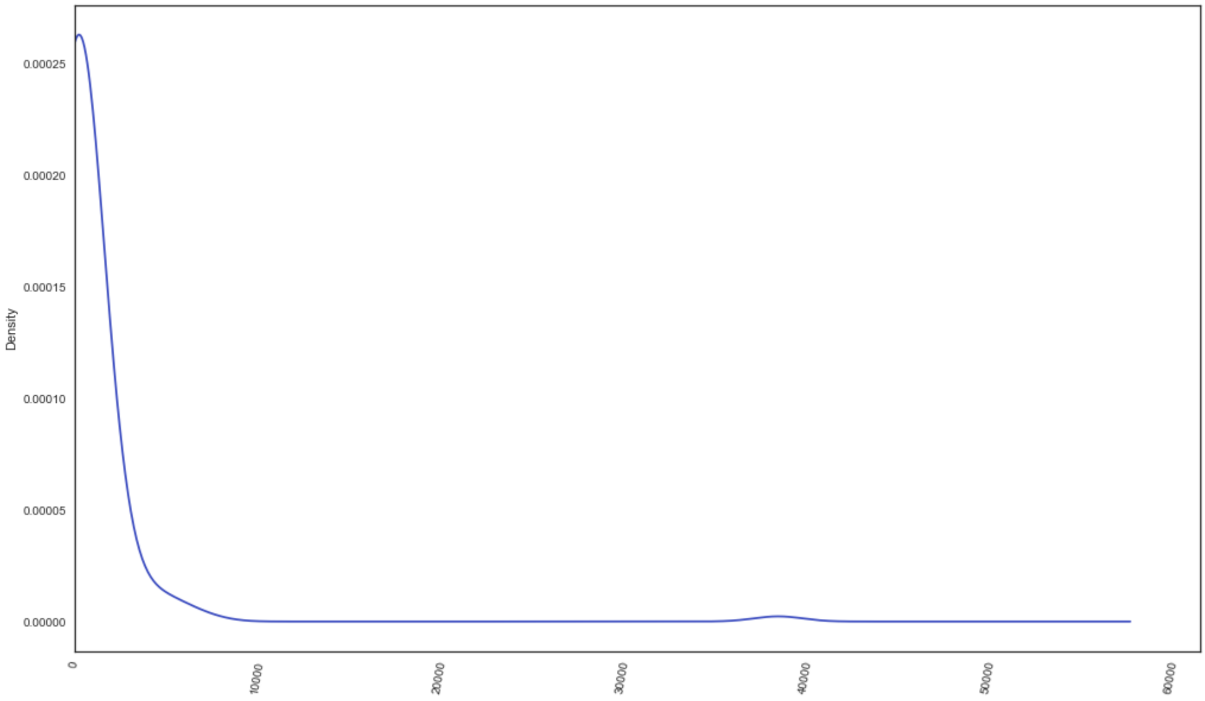


Figure 2 - Distribution of the Time variable

At this point the data was pickled and ready for the analytical steps.

B. Modeling

Method and evaluation

The first step of the analysis was to divide the model between a training, a validation and a test set. The first two sets are used to create a robust model and the last one is intended to evaluate its predictive power. The division between sets was done through scikit learn's train-test split where the model was split following a rule of thumb which separates the whole data set as $60-20-20\%$ for respectively the training, validation and test sets. At this stage, the test set was pickled and untouched until the most robust model was trained.

Before building a model, through data investigation was led. Table 1 provides a high-level view of the clean dataset through summary statistics.

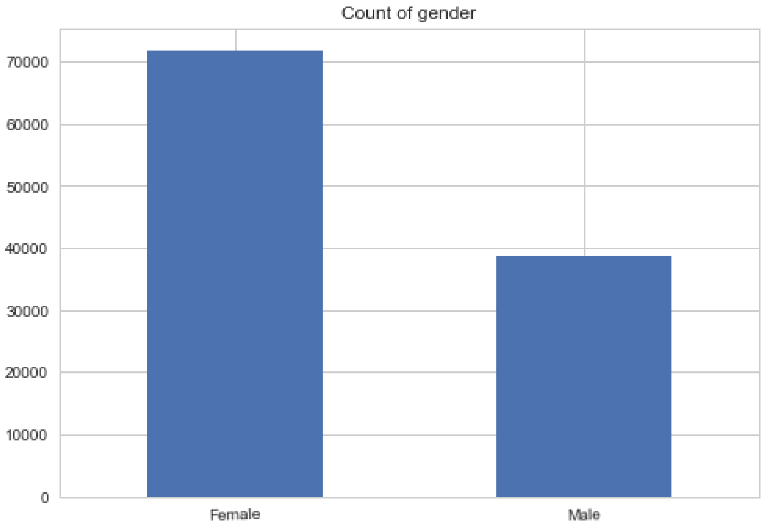
Number of Observations: 110 526

Variable	Min	Max	Mean	Std. Dev	Notes
Noshow	0	1	0.20	0.40	1 it the patient does not show up; 0 otherwise
Patient ID	N/A	N/A	N/A	N/A	Could help identify if no show has recurrence patterns
Gender	0	1	0.35	0.47	1 if male; 0 otherwise
Age	0	115	37	23.12	-
Scholarship	0	1	0.10	0.29	1 if true; 0 otherwise
Hypertension	0	1	0.20	0.39	1 if true; 0 otherwise
Diabetes	0	1	0.07	0.26	1 if true; 0 otherwise
Alcoholism	0	1	0.03	0.17	1 if true; 0 otherwise
Handicap	0	1	0.02	0.14	1 if true; 0 otherwise
Reminder	0	1	0.32	0.46	1 if true; 0 otherwise
Time	0	178	12.72	13.56	In days
Days	0	6	N/A	N/A	Where Monday = 0 and Sunday = 6
Neighbourhood	0	2	N/A	N/A	Groups 1,2 or 3 described in figure 1

Table 1 – Description and summary statistics of variables

Also, we summarize the data investigation that was led for main features:

Gender: By studying gender we are interested in whether the sex of the individual affects the probability of her showing up to a doctor appointment. In our data, as displayed on figure 3 there is large sample disproportion between men and women. This is in line with studies such as Tabenkin et al. (2004) that empirically demonstrate that women go to the doctor more often than men. Although this is not the focus of the study, interesting underlying patterns can be investigated by comparing age differences in gender doctor visits for instance-displayed in figure 4. From figure 4, although overall women tend to book doctor appointments more than men, this is not the case during childhood and teenage years where this proportion is similar.



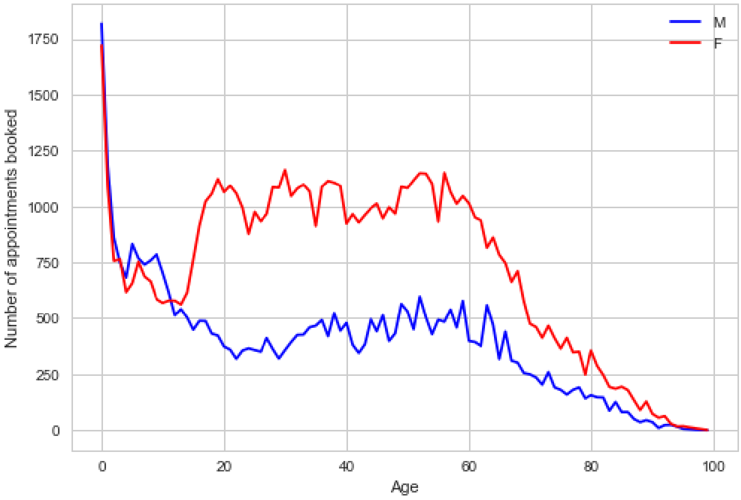


Figure 3 - Count of Women and Men

Figure 4 - Appointments by age and gender

Controls: Here, we include all the variables that act as controls. They are all summarized in table 1. We first note that most of the features, such as Handicap or Scholarship demonstrate low variance and must be interpreted with caution even if significance is drawn. This is because, in the sample, there are few individuals with handicaps. If a handicap is among patients who do not show up, it implies a high proportion of individuals with a handicap who do not show up, which is not accurate.

Timing: In this study, this set of variables in this section are of particular interest. The main questions that arise are:

- Is an individual more likely to not show up to an appointment when the date it is set at is distant from the date of the appointment?
- Does the day on which the appointment is set at influence whether individuals show up to the appointment?

The validity of these questions is further emphasized when exploring figures 5 and 6. Figure 5 demonstrates that no appointments are made on weekends (likely because clinics are closed then) and more appointments are scheduled for the beginning of the week then the end. Figure 6 suggests that as the difference between the appointment booking and its date increases, the more likely an individual is to break an appointment. It is noteworthy to mention that the average waiting time for patients between booking and appointment is about 13 days.

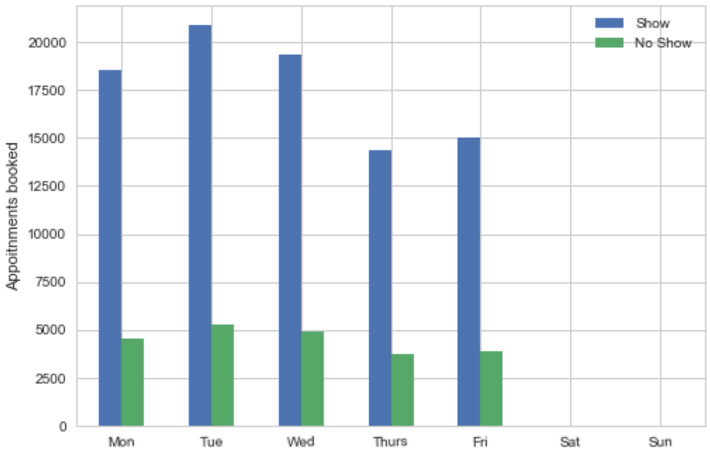


Figure 5 - Appointments by day of the week

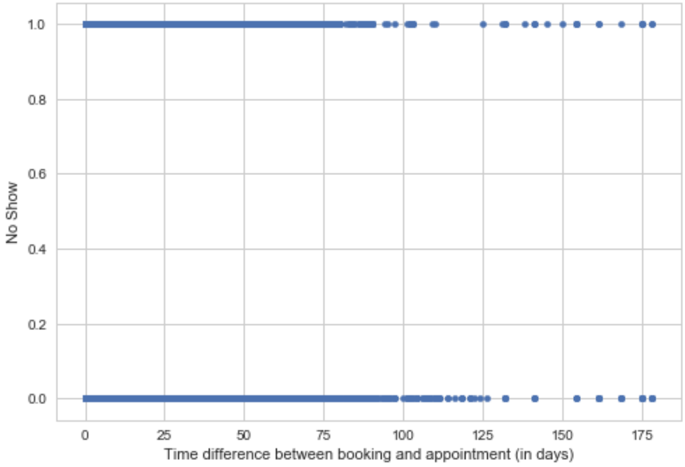


Figure 6 - Scatter plot of 'Time' vs 'No Show'

After the investigation of the features, we focus on building the model. As we are primarily interested in the questions brought forth in the above 'Timing' category, we build two models that attempt to shed light on these issues. The models are based on the following logistic regression:

$$\log\left(\frac{Noshow_i}{1-\ Noshow_i}\right) =\ \alpha + \gamma Timing_i + \beta_j Controls_{ij} + \varepsilon_i$$

In the first model (model 1), the Timing is the Time variable and in the second we use the Days feature.

To fit the model, two Python packages were evaluated: scikit learn's linear˙model and statsmodel (the Logit function). As the statsmodel was more flexible and clear in display of results, it was chosen here. Table 2 demonstrates the results of the models ran on the validation sets. Interesting patterns are brought forth by the logistic regression and discussed in depth in section III.

In this section, we take attention the p-values of the coefficients. As one of the goals of the study is to build a robust predictive model, we aim for the model with the highest accuracy while maximizing the degrees of freedom. In this case, we run the simple following tests:

- H0: $\beta_j \neq 0$ for all j (and $\gamma\ \neq 0$)
- H1: $\beta_j = 0$ for all j (and $\gamma = 0$)

Using the p-values, we conclude that the variables PatientID, Gender, Neighbourhood and Dayofweek do not have any statistical value at 5% significance. By discarding them, we created our third model.

To test how robust the models were, their prediction power was tested on the validation set and compared through the Mean Squared Error (MSE). As the fitted values are probabilities, to calculate the MSE a threshold at 0.5 was applied to the fitted valued of the model to create binary outcomes. Table 2 sums up the MSE of each model. As the three regressions seem to have approximately the same predictive power, the one with the highest numbers of degrees of freedom was retained as the best[2]. Other models with variables interactions were tested but this model remained the most robust due to insignificance of all additional variables. From thereon, the model could be tested in terms of predictive power on the test set.

[2] Although in this large sample, degrees of freedom are not a crucial concern, this choice made to find a theoretical support to compare models that perform very similarly.

	Model 1	Model 2	Model 3
Patient ID	-2.046e-17 (0.595)	-1.531e-17 (0.690)	-
Gender	-0.0373 (0.077)	-0.0393 (0.062)	-
Age	-0.0079* (0.000)	-0.0075* (0.000)	-0.0080* (0.000)
Scholarship	0.2364* (0.000)	0.2157* (0.000)	0.2426* (0.000)
Hypertension	0.0351 (0.270)	-0.0493 (0.120)	-0.0493* (0.000)
Diabetes	0.2400* (0.000)	0.2143* (0.000)	0.2218* (0.000)
Alcoholism	0.3146* (0.000)	0.2814* (0.000)	0.2966* (0.000)
Handicap	0.2933* (0.000)	0.2703* (0.000)	0.2838* (0.000)
Reminder	0.5491* (0.000)	0.6506* (0.000)	0.5503* (0.000)
Neighbourhood	0.0045 (0.776)	0.0163* (0.298)	-
Time	0.0126* (0.000)	-	0.0127* (0.000)
Days	-	-0.0042 (0.557)	-
MSE (On validation set)	0.2026	0.2027	0.2023

*Table 2 – Model Results with in parentheses p-values and where * represents significance at 5%*

Prediction

In this section we discuss the model’ s prediction performance. The model was fit on the test set and resulted in MSE = 0.2020, similar to the one found on the validation set. This low number shows that the model is accurate in its prediction.

To shed more light on the prediction power of the model, a confusion matrix was constructed using the scikit learn confusion˙matrix function coupled with the confusion matrix plot source code available on scikit learn – figure 7 This display shows that the model accurately predicts accurately most patients who do not show up to an appointment. Furthermore, in this specific case, the aim of the model should be to minimize the number of false positives (i.e.: the number of patients the model predicts will show but who do not) as the overarching aim is to minimize doctor idle time. In this respect, the model underperforms as there are more false negatives than false positives. To correct this, there are two methods that could be considered: (i) changing the threshold to create an asymmetric prediction (ii) a maximum score estimator. However, we do not implement them here as (i) does not provide more analytical insight and risks overfitting and (ii) is less efficient than the Maximum Likelihood Estimator - Kim et Pollard (1990). Overall, the model is satisfactory and provides evidence that the features evaluated are important predictors for whether a patient will show up or not to an appointment.

To create a richer model than the one at hand, there several possibilities: one could add features that measure whether the patient is punctual or not for instance; or one might consider modelling through categorical variables what type of medical professional the patient is seeking (e.g. it is likely that a parent going to a paediatrician for their child is less likely to break an appointment than a patient going to a general practice.)

MSE = 0.202

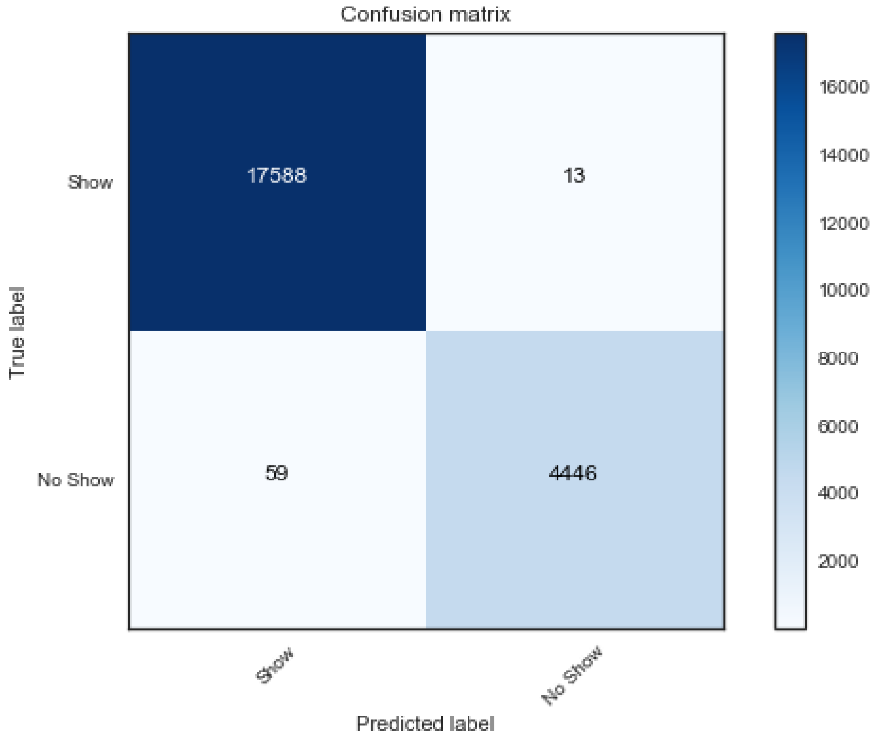


Figure 7 - Confusion matrix of the predictive model

III. Findings and Reflections

This section aims to discuss the results uncovered during the analytical process. First, we consider the coefficients that did not display any statistical significance. It is likely that certain of these variables would perform differently in other circumstances such as a more granular definition of features or a larger sample. For instance, the feature PatientID could shed light on the recurrence of appointment breaking by patients (i.e. If a patient breaks the appointment once she might be more likely to break again) if it was set in a larger or more suitable sample dataset (e.g. study over several years through a panel dataset). Neighbourhood was divided into a categorical variable of three groups. However this method creates a loss of information that is inherent to each neighbourhood. For example, adding the income, crime rate or even number of hospitals per neighbourhood could describe geographical locations better and uncover findings such as the ones of Mead (2017) who demonstrates a higher propensity of appointment breaks in rural areas.

It is interesting that the Gender did not show any significance. With this regard, the literature seems to display different effects. For instance, while Mead (2017) finds that women break appointments more than mean, Devasahay et al.(2017) finds converse results. It is difficult to draw a clear effect, which seems to suggest that there is no direct pattern linked to Gender.

Most of the controls variables were significant in our model and often displayed a positive effect on the probability of not showing up to an appointment. It is however important to bear in mind that due to the low variance of the variable interpretation must be done cautiously. However, we can affirm that in this model, if an individual has diabetes, it increases his probability of not showing up to an appointment by about 21%. Similar conclusions can be drawn for Scholarship, Alcoholism and Handicap. Due to the caution exercised in interpretation here, it is not reasonable to generalize such results. Furthermore, some results totally contradict the overwhelming evidence in the literature and should not be taken as accurate. Such is the case for Reminder, which according to the model leads to a large increase of no-shows[3].

In this context, Age is a commonly studied feature in the literature. Often it is considered that Age is a clear predictor to appointment breaks. This is what this study suggests as well as for each increase in the patient's age by a year, the model indicates a decrease in 0.8% probability of not showing up to an appointment. This is a common finding but contradicts recent research such as Devasahay et al. (2017) who find no significance of this variable.

This study also aimed to provide evidence, if any, that the moment of appointment booking influenced the no-show of patients. The findings of the model suggest that the booking day of the week bears no statistical significance on no-show. However, the time between booking and actual appointment is an important driver of appointment breaks. The model suggests that for each day between the booking and the appointment there is a 1% increase in probability of breaking the appointment. This finding is in line with empirical consensus (Deyo et Inui, 1980; Bean et Talaga, 1995; Grunebaum et al., 1996). Although this is a known result, it provides relevant insight to healthcare policy planners: it is in the interest of patients and doctors to set appointments as early as possible.

[3] The variable was closely inspected because of this anomaly but no evidence for the reason why has been found. One explanation is that the variable was not well defined by the author of the dataset and actually has 0 if the patient received a reminder and 1 otherwise.

References:

- Devasahay, S., Karpagam, S. and Ma, N. (2017). Predicting appointment misses in hospitals using data analytics. *mHealth*, 3, pp.12-12.
- Deyo, R. and Inui, T. (1980). Dropouts and Broken Appointments. *Medical Care*, 18(11), pp.1146-1157.
- Guven Uslu, P. (2007). Measuring Efficiency in Health Care: Analytic Techniques and Health Policy20071R. Jacobs, P.C. Smith and A. Street. Measuring Efficiency in Health Care: Analytic Techniques and Health Policy. Cambridge University Press, 2006. xvii + 243 pp., ISBN: 10-0-521-85144-0. *International Journal of Pharmaceutical and Healthcare Marketing*, 1(3), pp.264-265.
- McLean, S., Booth, A., Gee, M., Salway, S., Cobb, M., Bhanbhro, S. and Nancarrow, S. (2016). Appointment reminder systems are effective but not optimal: results of a systematic review and evidence synthesis employing realist principles. *Patient Preference and Adherence*, p.479.
- Mead, P. (2017). Understanding Appointment Breaking: Dissecting Structural Violence and Barriers to Healthcare Access at a Central Florida Community Health Center. *Graduate School at Scholar Commons*. [online] Available at: http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=8095&context=etd [Accessed 7 Dec. 2017].
- Bean and Talage (1995). Predicting Appointment Breaking. Journal of health care marketing, 15, pp.29- 34.Physiotherapy, 101, pp.e980-e981.
- Anderson, K. (2014). An overlapping appointment scheduling model with stochastic service time in an outpatient clinic. Operations Research for Health Care, 4.
- Ahmadi-Javid, A., Jalali, Z. and Klassen, K. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. European Journal of Operational Research, 258(1), pp.3-34.
- Deceuninck, M., Fiems, D. and De Vuyst, S. (2017). Outpatient scheduling with unpunctual patients and no-shows. European Journal of Operational Research, 265(1), pp.195-207.
- Kim, J. and Pollard, D. (1990). Cube Root Asymptotics. The Annals of Statistics, 18(1), pp.191-219.
- Tabenkin, H., Goodwin, M., Zyzanski, S., Stange, K. and Medalie, J. (2004). Gender Differences in Time Spent during Direct Observation of Doctor-Patient Encounters. Journal of Women's Health, [online] 13(3), pp.341-349. Available at: http://online.liebertpub.com/doi/abs/10.1089/154099904323016509.