Predicting Medical Appointment No-shows

Knut Ulsrud

6/4/2020

Overview

This section presents the project's goals and a high-level overview of project data.

Goal of the project and key steps that were performed

This project's goal was to predict attendance (i. e. No-show vs. attended) for doctor's appointments based on publicly available medical appointment data. This report was produced as a submission to the Data Science: Capstone course provided by HarvardX on the EdX platform.

Dataset Description

The medical appointment dataset has 110,527 rows and 14 columns, and contains patient attendance data from April 29 to June 8, 2016, for a clinic in Brazil. The original dataset was sourced via Kaggle. https://www.kaggle.com/joniarroba/noshowappointments. Columns contain the following information:

Variable name	Variable Description
1. PatientId	Unique identifier for each patient
2. AppointmentID	Unique identifier for each appointment
3. Gender	Male or Female
4. ScheduledDay	When the appointment was scheduled
5. AppointmentDay	When the appointment happened
6. Age	Patient age
7. Neighbourhood	What neighbourhood patient lived in
8. Scholarship	Binary variable related to scholarship reception through the Bolsa Familia program
9. Hypertension	Whether the patient lived with hypertension
10. Diabetes	Whether the patient lived with diabetes
11. Alcoholism	Whether the patient lived with alcoholism
12. Handicap	Whether the patient lived with a physical handicap (i. e. blindness, inability to walk, etc.)

13. SMS_received Whether the patient received an SMS reminder

14. No_show Whether the patient showed up for their appointment

The following table shows the first four rows of the dataset.

PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No_show
2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No

Methods and Analysis

This section presents steps taken to perform initial analysis on the dataset. It includes the following sections: Data Preaparation, Data Exploration and Visualization, Insights Gained, and Modelling Approach.

Data Preparation

A high-level check revealed that there were no missing values in the appointment dataset. The data was already presented in a tidy format (one row equals one observation / appointment) which means no additional transformation was required to start using the dataset.

There are two date columns which were provided in year-month-date – hour-minute-second format (i. e. “2016-04-29T00:00:00Z”). To facilitate analysis, both ScheduledDay and AppointmentDay columns were transformed into the following:

- Day of week
- Hour of day (only available for ScheduledDay)
- Duration from booking to appointment

For AppointmentDay all hour-minute-second data was set to 00:00:00. This means that calculated duration from booking to appointment might err by the number of hours from 00:00:00 to the actual appointment time.

Data Exploration and Visualization

It is likely that there is systematic variation between patient specific information and whether or not they No-show. This section will explore this variation according to

- Scheduling and appointment timing
- Past no-shows
- Demographics effects (i. e. age, gender, etc.)
- Medical information
- Other binary identifiers

It will also explore basic characteristics of variables of interest.

As shown in the table, roughly 20% of appointments in the dataset were missed.

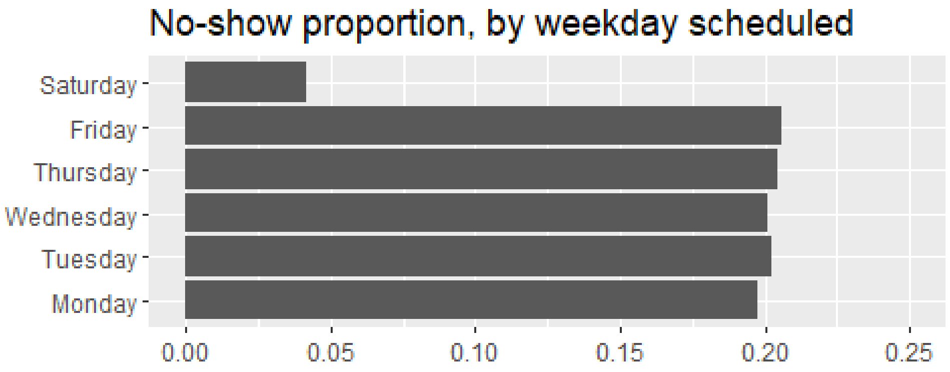
No-show	Proportion
FALSE	0.798
TRUE	0.202

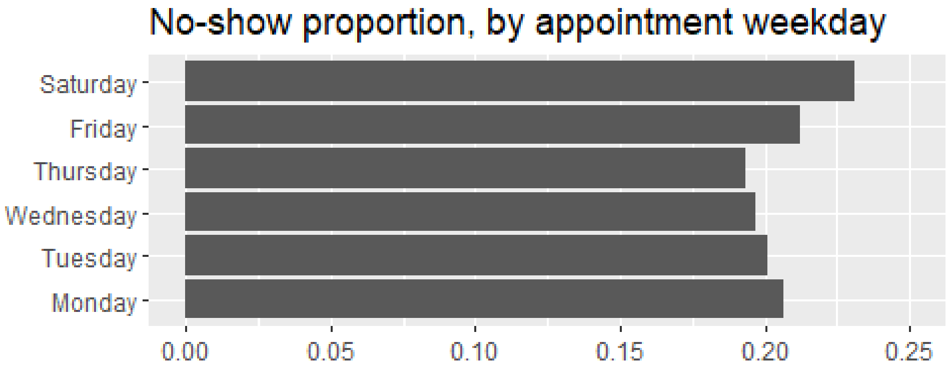
Appointment Timing

This section explores variation between time-related variables and appointment No-shows.

Appointment and scheduling Weekday

At first glance it appears that appointments that were scheduled on Saturdays were much less likely to be missed. Otherwise, scheduling day does not appear to vary by meaningful proportions.





On the other hand, appointents held on Saturdays appear to be missed more often than other days, while Thursday appointments are kept most often. There is minor variation between other days of the week.

Further exploration reveal that that much fewer appointments were scheduled and held on Saturdays than on other days. This indicates that Saturday data may be less reliable than other weekdays, as it may be susceptible to for instance selection bias.

Scheduled Weekday	Appointments	Percent No-show
Monday	23085	0.20
Tuesday	26168	0.20
Wednesday	24262	0.20
Thursday	18073	0.20
Friday	18915	0.21
Saturday	24	0.04

Appointment Weekday	Appointments	Percent No-show
Monday	22715	0.21
Tuesday	25640	0.20
Wednesday	25867	0.20
Thursday	17247	0.19
Friday	19019	0.21

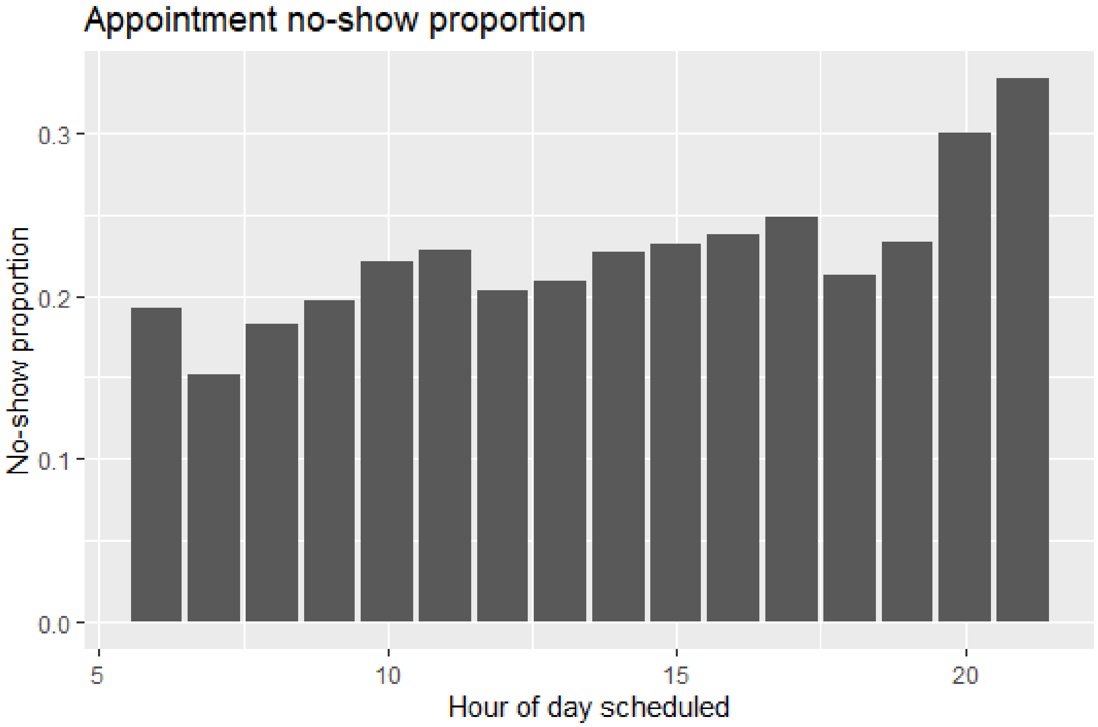
Saturday	39	0.23

Instead of filtering for Saturday appointments and removing useful data, a more prudent decision would be to drop Weekday Scheduled.

Saturday appointment effects are small enough that it should not meaningfully skew the data, and the variable's importance to the final model could be tested. This would confirm whether the variation between appointment weekdays is explained by other variables.

Hour of day scheduled

The hour of day that appointments were scheduled also seems to have an impact, varying from around 15% No-shows early in the morning, to over 30% late in the evening. It is unclear why this should be the case, other than that patients may be more tired and forgetful and therefore forget appointments scheduled later in the day.

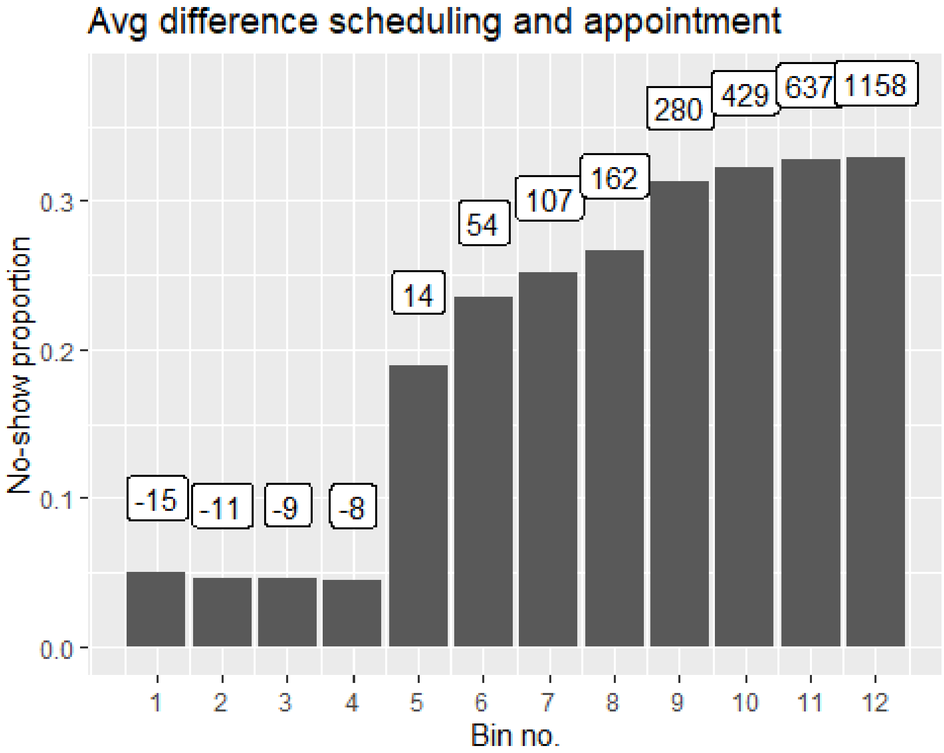


Difference between scheduling and appointment

By binning the difference between scheduled time and appointment time we see that there is substantial and systematic variation between time differences, and No-shows.

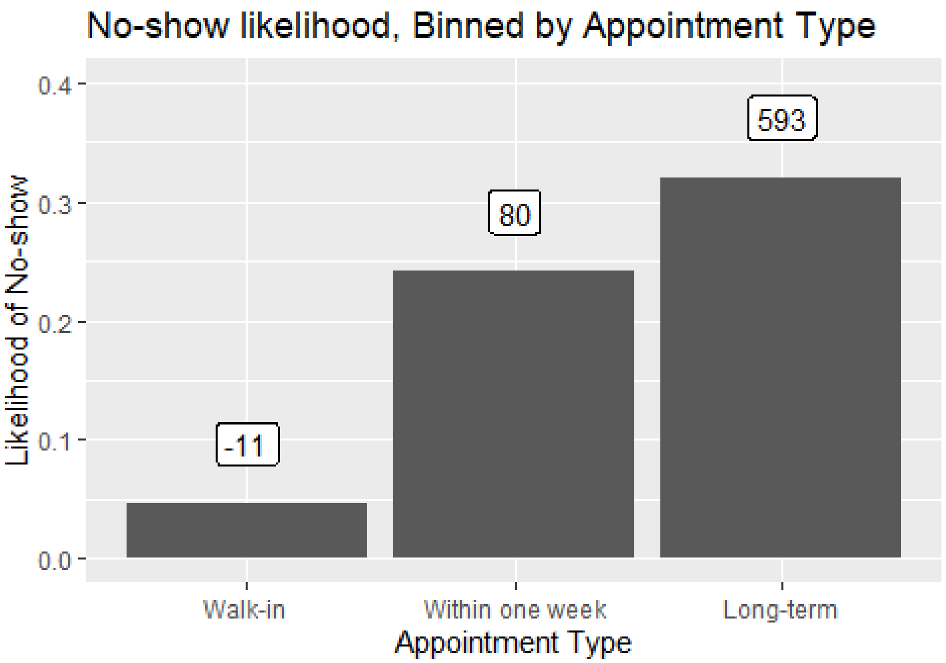
Interestingly there are appointments that are recorded as scheduled after the appointment took place. This has two potential explanations. One is the aforementioned issue where exact appointment time was not reported, only date and 00:00:00 for hour-minute-second. This would mean that any same-day scheduled appointment would appear negative. Second, a patient walk-in could happen, and entered into the system by an administrator later.

Either way, the recorded difference between scheduling and appointment timing deviates from the true difference by a number of hours. This is made clear by the presence of negative values. This paper continues the timing difference exploration with this caveat in mind.



To do a high-level exploration of No-show outcomes it is reasonable to bin the hours difference data. This is especially the case given that individual observations might vary within a few hours from their actual time, as just explained. This would smooth data recording errors while retaining high-level variation.

Binning was based on the intuition that three common appointment types vary by Walk-ins (less than 1 hours difference), Scheduled within a week (1 to 168 hours difference), and Scheduled within more than one week (more than 168 hours difference)



Binning shows that the intuition about appointment types is reasonable for the purposes of exploratory analysis.

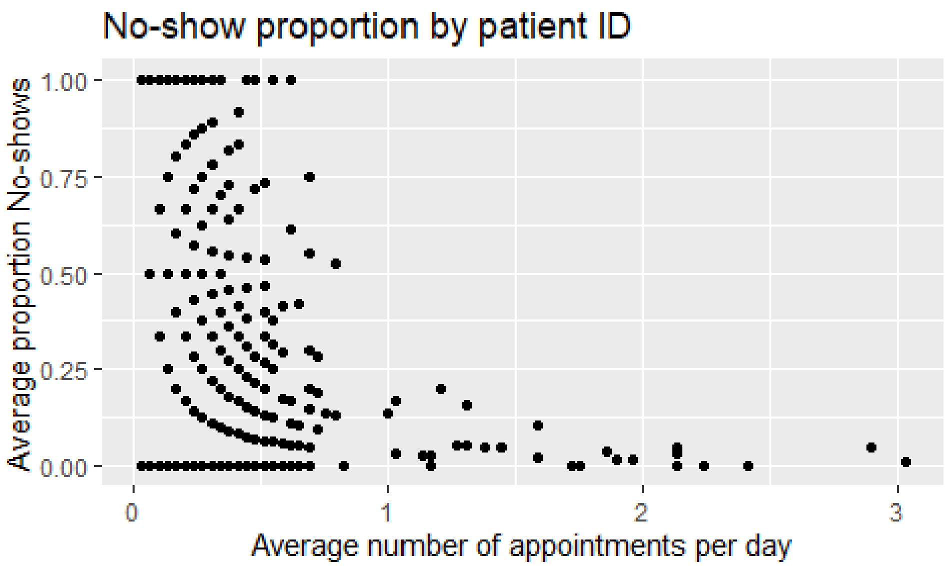
Demographics

This section explores variations in the No-show outcome by different demographic factors:

- Patient effects
- Age
- Gender
- City Neighbourhood patient lives in

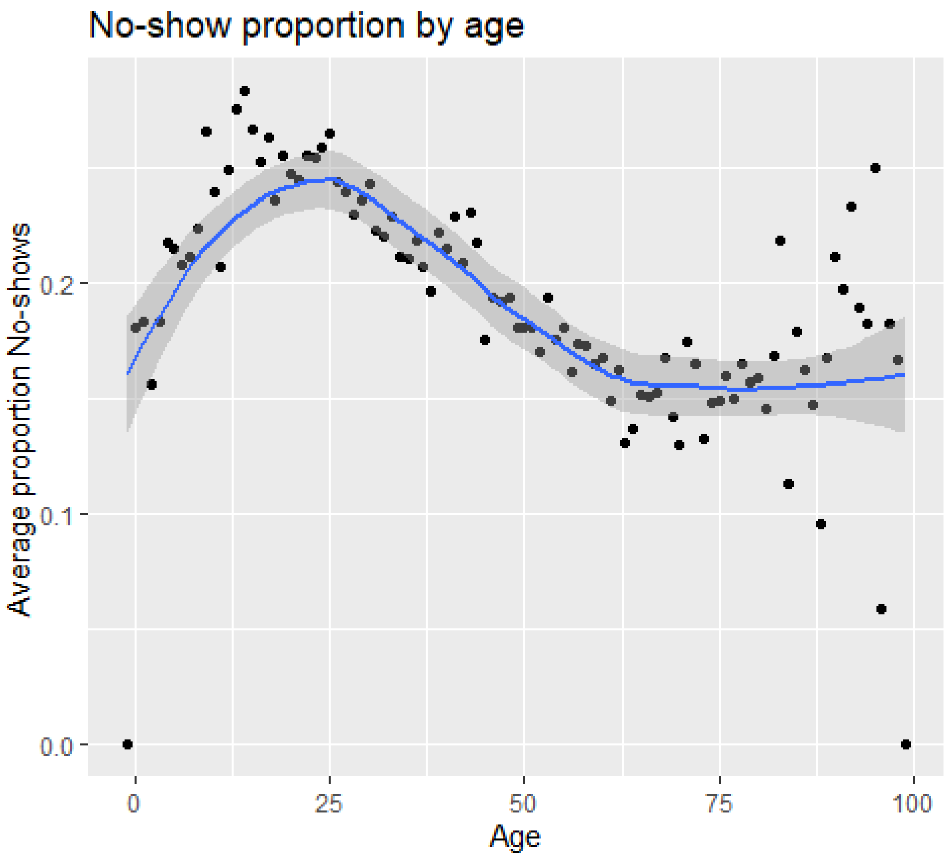
Patient effects

The scatterplot confirms the intution that there are differences in how many appointments patients had over the period of the dataset, and how likely they were to miss their apointments. For instance, there is a sharp cutoff in No-show likelihood around an average of one appointment per day. In other words, patients who see their doctor more frequently are more likely to keep their appointments. The observation that some patients have more than one appointment per day is reasonable. This is because one longer appointment would likely be scheduled as a sequence of individual appointments in medical scheduling software, giving the appearance of multiple appointments.



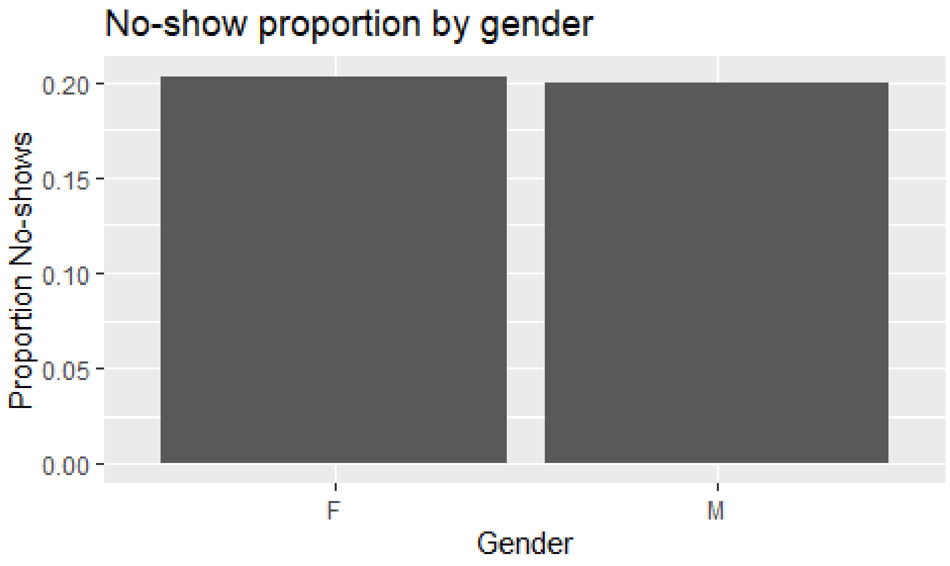
Age

There is a clear visual relationship between age and likelihood of no-shows. Patients in their 20s seem to miss appointments more often than other age groups, followed by gradually lower chance of No-show as patients age. Older patients diverge from this pattern, where some patients are reliable, and others not.



Gender

There does not appear to be any obvious connections between gender and likelihood of No-shows.



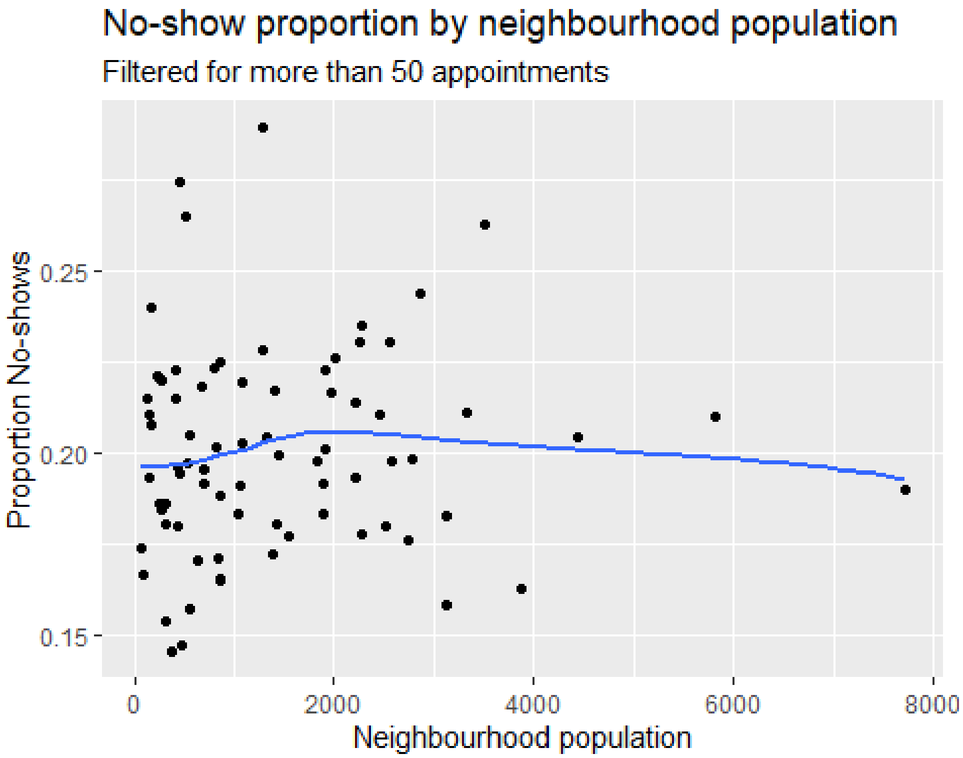
Neighbourhoods

Some neighbourhoods have high likelihood of misesd appointments, and others lower. However, because there are a limited number observations for some neighbourhoods, the difference in likelihood might be explained by other factors, or simply be person-dependent. This was confirmed by considering the number of people per neighbourhood, sorted by the top and bottom likelihood No-show neighbourhoods. Both high and low likelihood no-show observations tend to have few observations, with minor exceptions. For reference, mean observations per neighourhood is 1364.5308642

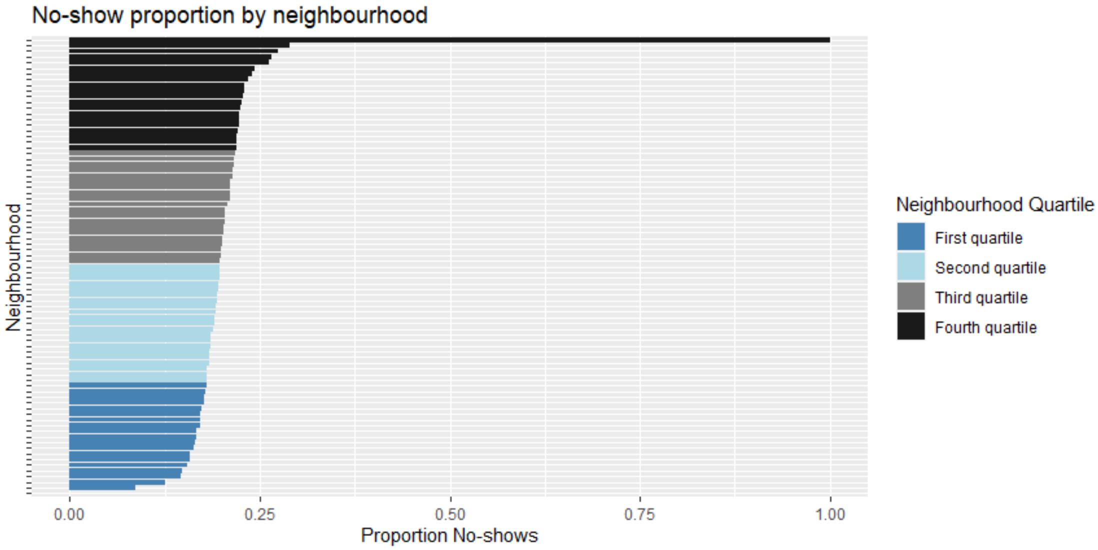
Neighbourhood	No_show (High)	n
ILHAS OCEÃ,NICAS DE TRINDADE	1.000	2
SANTOS DUMONT	0.289	1276
SANTA CECÃ• LIA	0.275	448
SANTA CLARA	0.265	506
ITARARÃ‰	0.263	3514

Neighbourhood	No_show (Low)	n
PARQUE INDUSTRIAL	0.000	1
ILHA DO BOI	0.086	35
AEROPORTO	0.125	8
MÃ• RIO CYPRESTE	0.146	371
SOLON BORGES	0.147	469

The scatterplot confirms that there is additional variation between neighbourhoods, that is not explained by the number of appointments from that neighbourhood. A filter for more than 50 neighbourhoods was applied in the graph to account for an outlier neighbourhood with 100% No-shows.



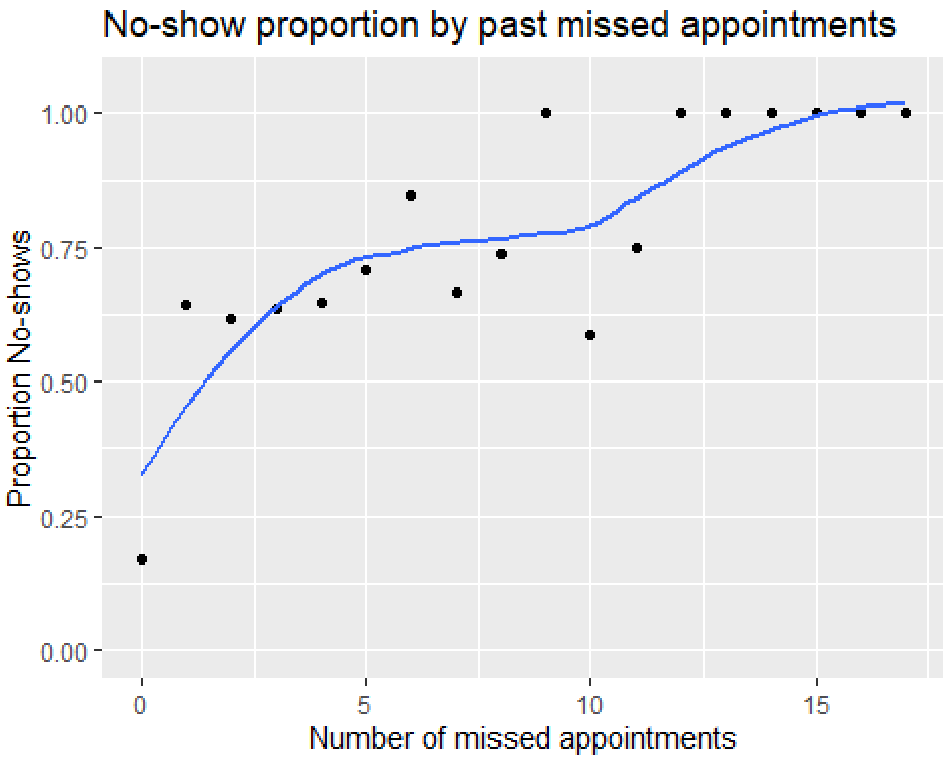
Neighbourhoods were subsequently grouped into quartiles, as shown in the graph. Neighbourhood names were taken out fo ease of presentation.



Previously missed appointments

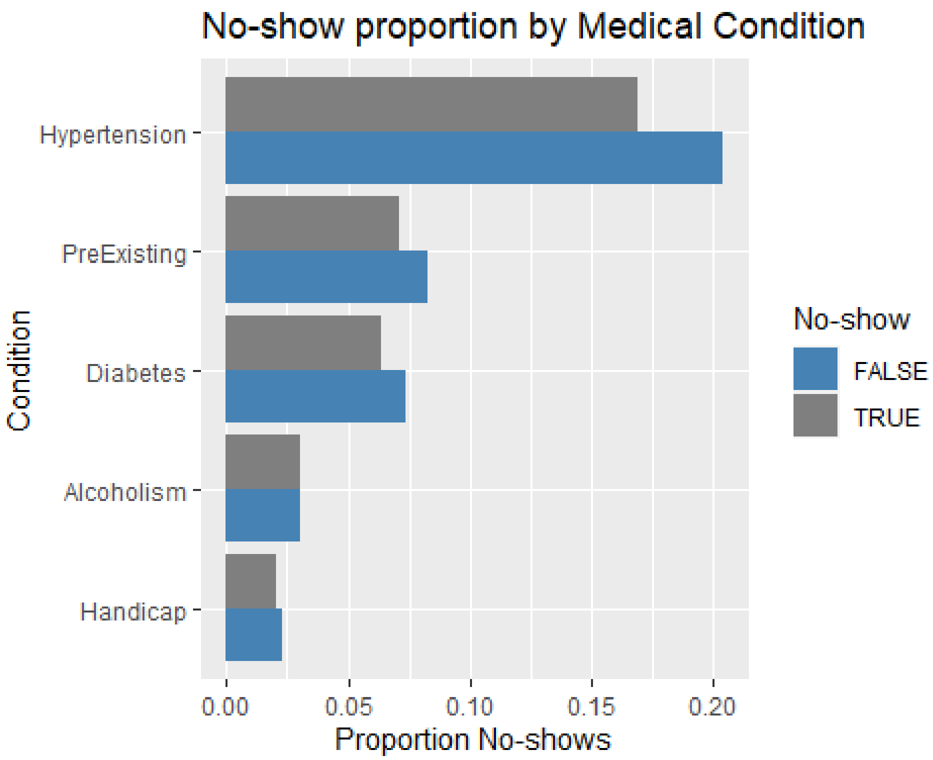
Intuition suggests that someone who have missed a past appointment may be more likely to miss subsequent appointments. This is confirmed by the data, which shows that patients are progressively more likely to miss appointments, the more appointments they have alreay missed.

There is a limitation in this as all the dataset misses a total patient history. This means that all individuals will start at no past no-shows, and accumulate them over the timeframe of the dataset.



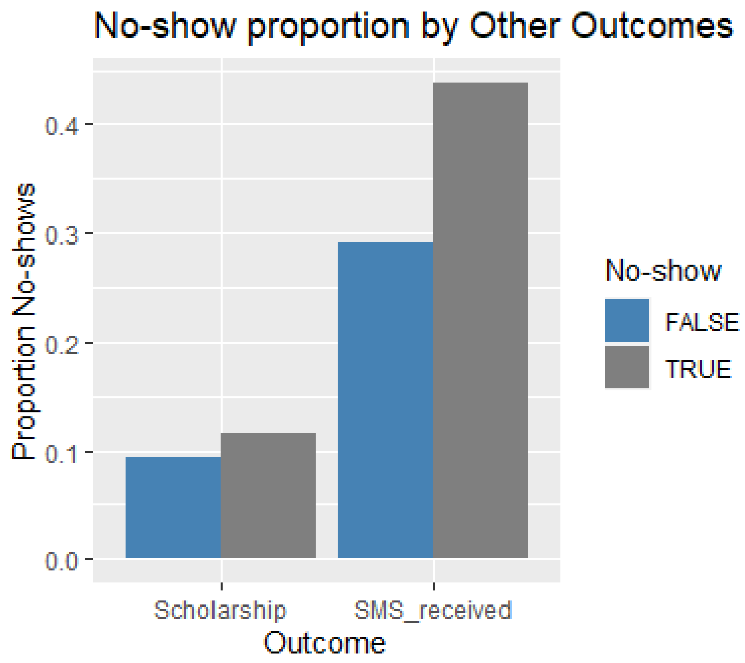
Medical Information

Medical information includes clinical diagnoses that were captured in the dataset. PreExisting is a variable adding up all conditions, to see if there are additive effects on likelihood of no-shows from having more than one medical condition. Individuals with hypertension were less likely to miss an appointment, as were those with Diabetes, however the latter has a minor effect only. The PreExisting variable does not appear to explain variation beyond what is already captured by the diabetes variable. Furthermore, as different medical conditions appear to influence the likelihood of no-shows in different directions, an additive variable is not likely to add value. Finally, alcoholism and Handicap shows minor to no variations with likelihood of no-shows.

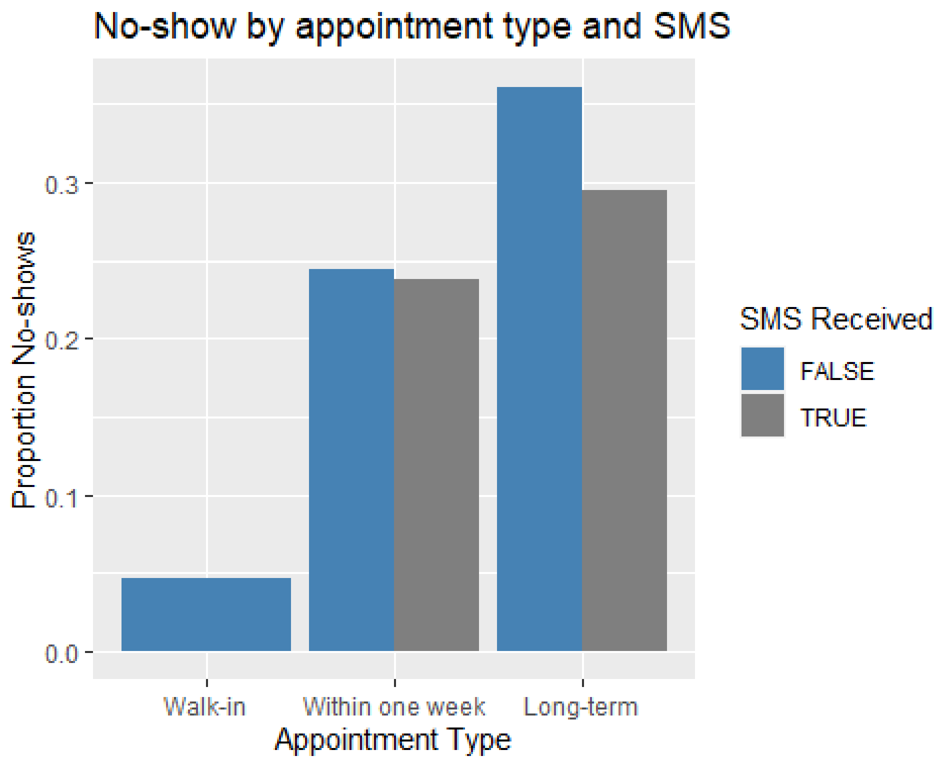


Other variables

Having received a Scholarship (related to the Bolsa Familia program) shows minor variation only. Having receied an SMS appears to make individuals less likely to make their appointment, which is counterintuitive. This surprising result warrants additional exploration to understand whether the data is reliable, or is inadvertently capturing relationships between other variables and the likelihood of no-shows.



It was identified that by analyzing the relationship between SMS received and likelihood of no-shows by appointment type that the correlation between SMS received and No-shows was explained by appointment type, defined by the duration from booking to the appointment. The graph shows that only medium to long-term bookings receive the SMS, and those appointment types (especially longer-term) were already shown to have a higher likelihood of no-showing. In fact, individiuals with long-term bookings are more likely to show up to their appointment if the recieved an SMS reminder.



Insights gained

Based on the explorative analysis the following changes were made to the dataset:

Taking out variables reduces noise and increases the likelihood that the machine learning models identify signals present in the dataset.

Take out:

- Unique identifiers, including PatientId, AppointmentId, and Date variables
- Neighbourhood is replaced by binned variables (quartiles)
- Handicap, Alcoholism, and gender, as they showed limited variation with No-shows
- Filter one patient that was recorded as 115 years old (outlier)
- Drop bins for appointment types and keep continuous scale

Add new variables:

- Number of appointments per day
- Neighbourhood bins
- Cumulative missed appointments

Feature Engineering

To facilitate algorithm analysis, the following feature engineering was completed:

- Categorical features transformed to factors

  - Appintment week day
  - Neighbourhood bin
  - No-show
- Numerical features centered and scaled
  - Age
  - Scheduled Hour
  - Past No-shows
  - Avg. Appointments Per Day
  - Difference between Booking and Appointment
- Binary features left as is
  - Scholarship
  - Hypertension
  - Diabetes
  - SMS_received

A logical test was confirmed that all features had been accounted for.

Modeling approach

The goal of the project is to predict whether an appointment will be a No-show or not, which means the problem is well suited to classification models. Roughly 20% of apointments are No-shows, which means that the dataset is imbalanced. All models predict the probability of No-show, which enable optimizing the probability threshold towards higher certainty but fewer correct predictions, against lower certainty but more correct predictions.

The project followed the following approach to evaluating models:

1. Use a basic classification algorithm with low processing time for preliminary results, and to explore strategies for imbalance mitigation.
2. Visualize results to better understand which features are most important for predicting No-shows.
3. Using only the most important features and a reduced dataset to reduce processing time, test a range of more sophisticated models and rank their results.
4. Run top performing models using the full dataset, and tune parameters for to optimize predictions.
5. Optimize the probability threshold for the top models.
6. Make final prediction on the test set using the one best performing model from step 5.

The data was split into 20 / 80 proportions, where 20% is set aside as the final test set, and 80% is used to train the model(s). Cross-validated versions of the training set is used throughout to evaluate models, until the final prediction on the test set in step 6. 20% was selected as convention, leaving enough data to train in the train set, while keeping enough data in the test set to appropriately evaluate the final trained model.

Evaluating models

Accuracy is a common metric to evaluate classification models, measuring the proportion of correct predictions. However, because it is based on sensitivity (true positives) and specificity (true negatives) it is biased for unbalanced datasets. In an extreme situation,the metric creates an incentive for the model to predict all appointments as negatives (i. e. not No-show), with accuracy remaining at 80%.

For this reason the project will use Precision and Recall (Sensitivity) to evaluate models' performance. Precision is used instead of Specificity as it optimizes for the proportion of true positives to all positives (i. e. how many predicted No-shows were actually No-shows). Sensitivity measures what proportion of all No-shows were predicted as No-shows.

Precision and Sensitivity generally move in the opposite direction. For instance, if Sensitivity is higher and more No-shows were captured overall, Precision is likely to be lower because this usually means there are more false positives. Inversely, when Sensitivity is lower and fewer No-shows were captured overall, Precision is likely to be higher because the model predicts No-shows for more certain observations. The trade-off between Precision and Sensitivity can be optimized using probability threshold optimization.

For the process of training and comparing different models, ROC will be used as it enables an easier comparison (one number vs. two in the case of Sensitivity and Precision). ROC will be calculated using cross-validation of training data.

Results

This section presents the results from the 6-step approach described previously.

1. Rpart Exploration

The first model uses Rpart as it is a fast-running classification model that can set a baseline to compare other models to.

The initial model produced an ROC of 0.578. An ROC of .50 is the equivalent of random guessing, meaning that there is a lot of potential improvement to be made. The maximum ROC value is 1, meaning the model would predict all values perfectly. All reported ROC values represent the average model performance on the cross-validated training set.

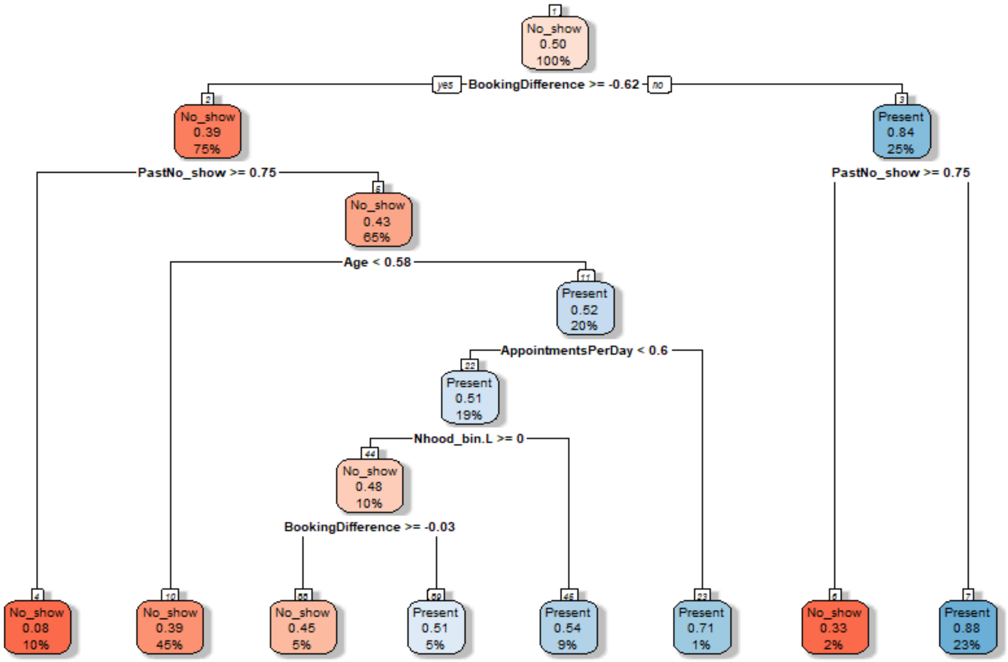
ROC	model
0.578	rpart - no sampling

Using imbalance strategies help improve results, taking the best model up to roughly 0.7 ROC. Results from different sampling strategies were similar, however smote was selected for being marginally better, as well as having the quality of creating synthetic data similar to the original data, as opposed to reusing or dropping data.

ROC	model
0.699	rpart - smote
0.699	rpart - down
0.698	rpart - up
0.698	rpart - rose
0.578	rpart - no sampling

2. Visualizing Results, and Variable Importance

Rpart is a tree model, which makes visualizing results easy. The preliminary results indicate that past No-shows are important for predicting future no-shows. Furthermore, age is also important, where high ages are likely to show up, and lower ages less likely to show up. (having seen the data this is an oversimplification, but it holds true directionally). Appointments that were far from their booking date are more likely to be missed. Finally, patients with more frequent appoitments are less likely to miss their appointments.



Using results from the preliminary analysis we can estimate variable importance for the different features. This enables us the next step of the analysis, which is to evaluate different models using a simpler version of the dataset to reduce processing time.

The calculated variable importance adds additional nuance to which features are considered important by the model. Importance is estimated on a scale from 0 to 100. In addition to the aforementioned, time of booking, SMS reminder, neighbourhood, and the hypertension variable are also considered marginally important by the model.

However, looking at the detailed importance values, having booked appointments on Mondays and living in certain neigbourhoods are only ewakly important. an arbitrary cutoff was set at importance of 2, where those below 2 was dropped. Note that unimportant features are only taken out for the model testing phase, and will be put back in for final model training.

Feature	Importance
BookingDifference	100.00
PastNo_show	85.42
Age	19.02
SMS_received	13.94
AppointmentsPerDay	7.66
ScheduledHour	3.66
Nhood_bin.L	3.51
Hypertension	3.44
Nhood_bin.Q	0.74
AppointmentWeekDayMonday	0.32
Scholarship	0.15
Diabetes	0.00
AppointmentWeekDaySaturday	0.00
AppointmentWeekDayThursday	0.00
AppointmentWeekDayTuesday	0.00
AppointmentWeekDayWednesday	0.00
Nhood_bin.C	0.00

A sense check confirms that little to no predictive power is lost by taking out the unimportant features.

ROC	model
0.699	rpart - smote
0.699	rpart - down
0.698	rpart - up
0.698	rpart - rose
0.578	rpart - no sampling
0.578	rpart - no sampling - important vars

3. Testing different models

A number of algorithms were tested on this classification problem. They include but are not limited to boosted trees, k-nearest neighbour, neural net algorithms. To reduce processing time, these algorithms were run on the first 10,000 rows from the training set, and only inlcuding what features were considered to be important, as specified previously. They were also run without smote-sampling for now.

Using different algorithms improved results beyond what was observed previously, even without using a sampling strategy. Of the 10 tested algorithms, 7 performed better than the smote-sampling rpart model that was trained on the full training set.

model	ROC
gbm - no sampling - important vars	0.759
mlp - no sampling - important vars	0.742
monmlp - no sampling - important vars	0.738
naive_bayes - no sampling - important vars	0.711
multinom - no sampling - important vars	0.705
glm - no sampling - important vars	0.704
avNNet - no sampling - important vars	0.700
rpart - smote	0.699

4. Run and tune top performing models on full training set

The top model was gbm - Generalized Boosted Regression Model. This is a boosted tree algorithm, a class of algorithms which is well suited to classification. The second model is mlp - Multilayer Perceptron. This is a neural net type algorithm.

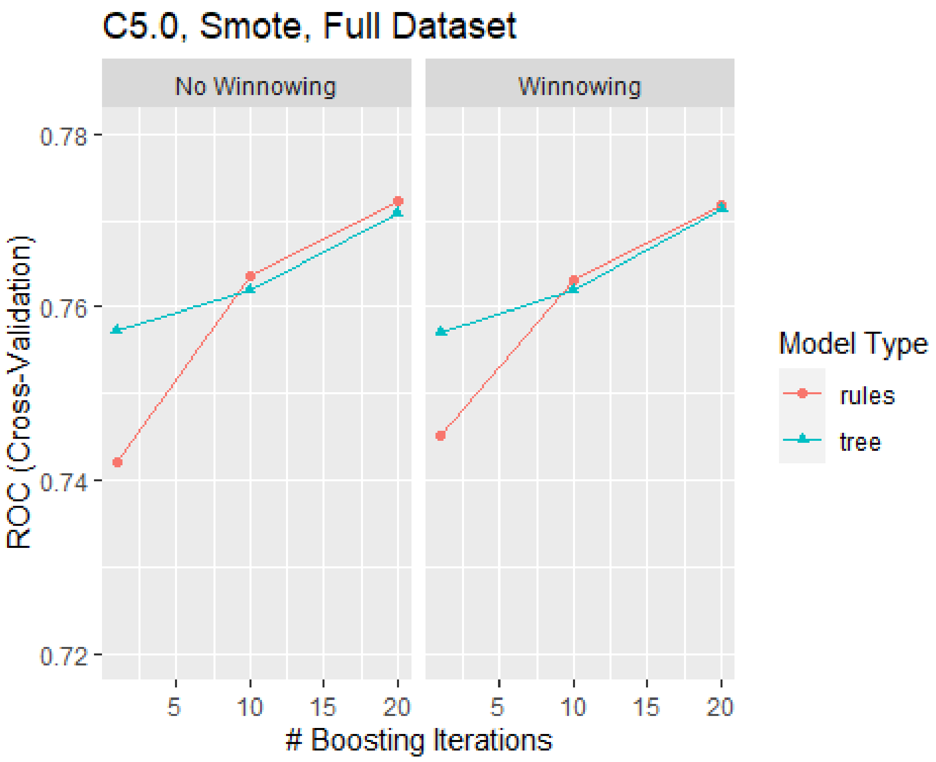
While not scoring highly in this instance, the author has had previous success using C5.0 on large datasets. C5.0's success with larger dataset would not be reflected in the 10,000 rows evaluated previously. Therefore, C5.0 will be the third model to be evaluated with smote-sampling and the full dataset.

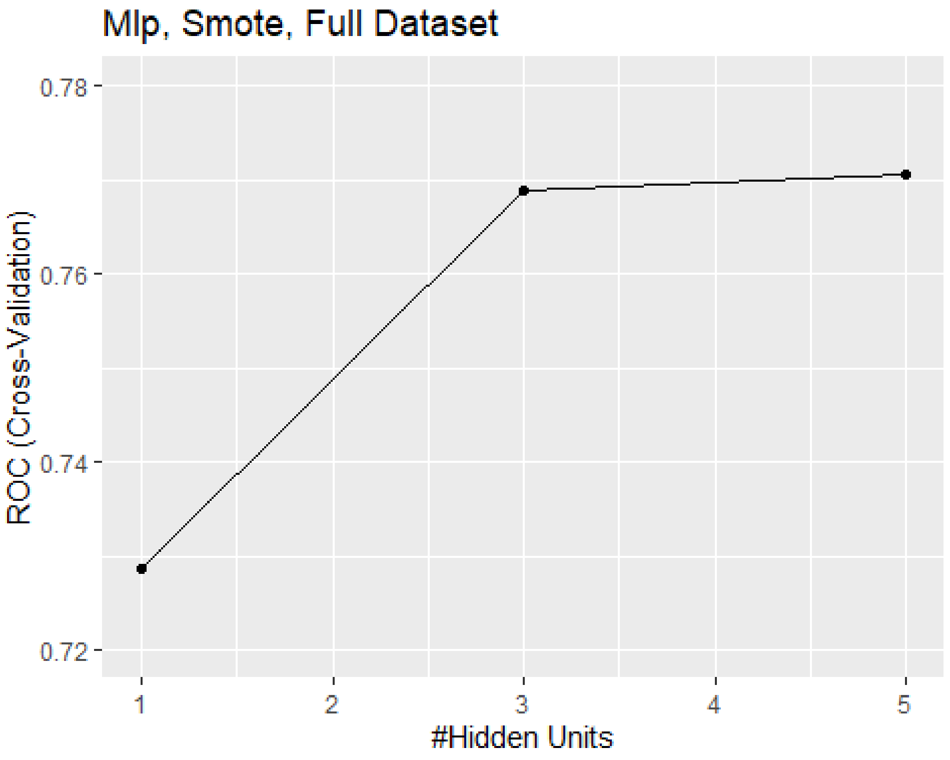
Results show that adding the full dataset improved results marginally for all three algorithms, with gbm performing slightly better than C5.0, followed by mlp.

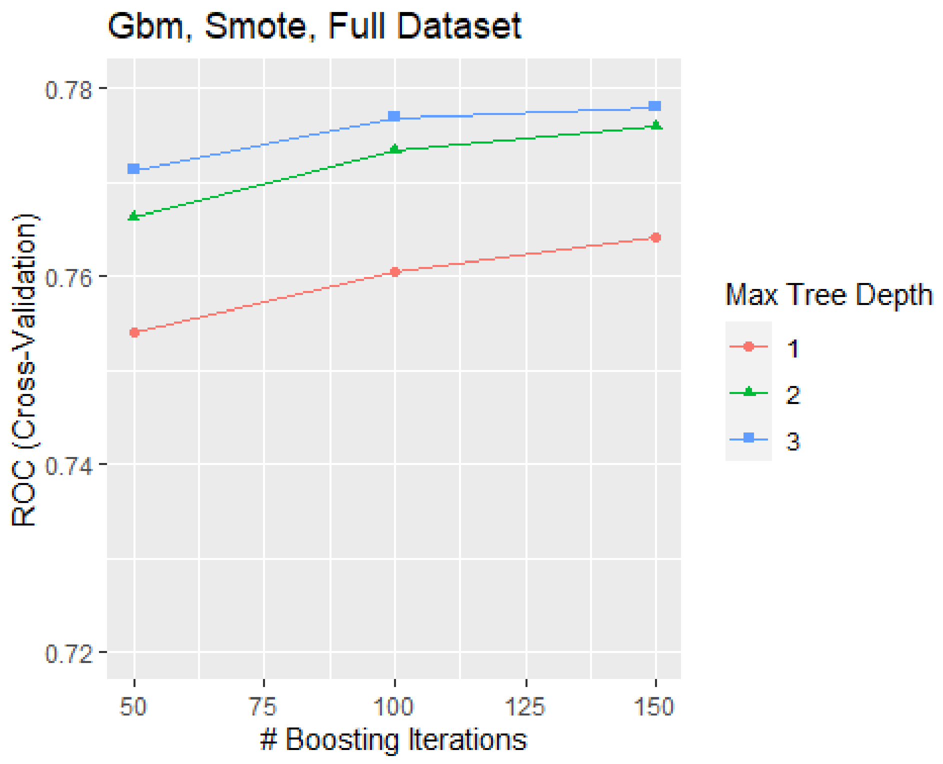
model	ROC
gbm - smote - full dataset	0.769
C5.0 - smote - full dataset	0.762
gbm - no sampling - important vars	0.759
mlp - smote - full dataset	0.756
mlp - no sampling - important vars	0.742

In considering each model's tuning parameters, C5.0 appears to have a a steeper curve than the other two models. This indicates that the potential for model tuning is higher for C5.0 than for the two other models.

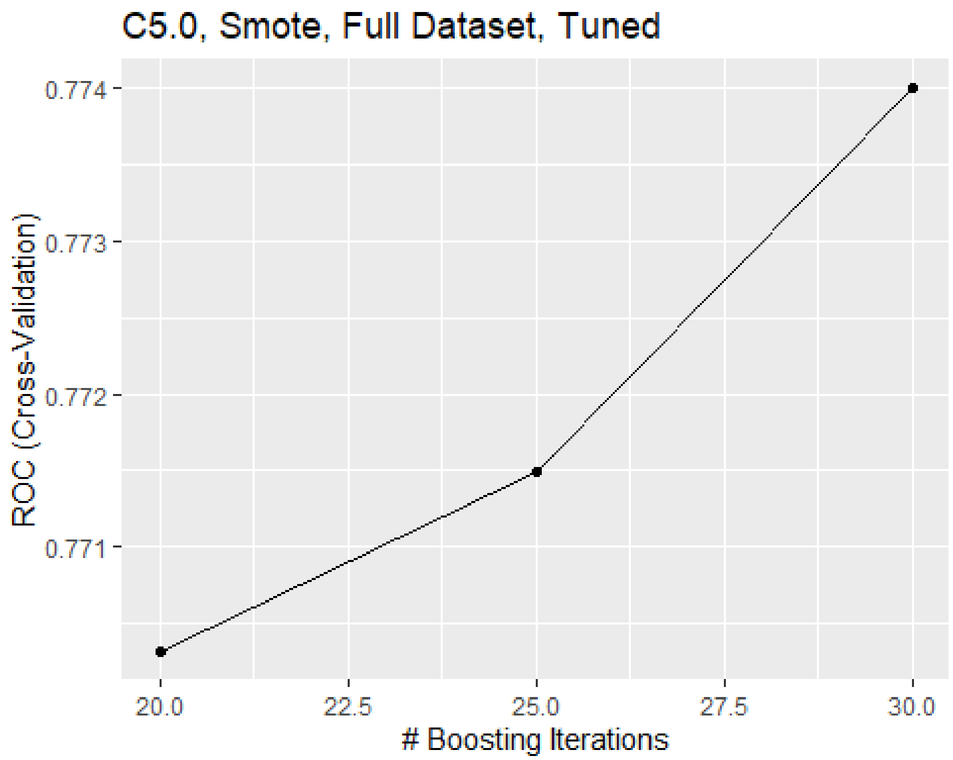
As a reminder, ROC values reported in the results tables are averages from the models' cross-validated training set. This is why values appear to be higher in the graphs.


C5.0, Smote, Full Dataset


Mlp, Smote, Full Dataset



For C5.0 winnowing apperas not to make a difference, with tree and rule based models both perform well. However, higher number of boosts seem to improve performance. The final model was trained setting “trials” to 20, 25, and 30. 30 trials was the highest R was able to run without crashing.



While the graph indicates additional potential gain, 30 iterations was the highest number that R could run without crashing. With the additional tuning, C5.0 is now the leading model.

model	ROC
C5.0 - smote - full dataset - tuning	0.772
gbm - smote - full dataset	0.769
C5.0 - smote - full dataset	0.762
gbm - no sampling - important vars	0.759
mlp - smote - full dataset	0.756

5. Optimize probability threshold for top models

The probability threshold (i. e. the threshold for when to predict a No-show or not on a scale from 0 to 1 predicted likelihood) can be optimized by using Thresholder. This function uses cross-validated training data to identify the probability threshold which optimize the trade-off between Sensitivity and Precision.

Precision was set to be at least 0.8, with otherwise maximized sensitivity. This was chosen because it is arguably more valuable to a doctor’s office to have fewer, but quite certain No-show predictions, than many uncertain ones.

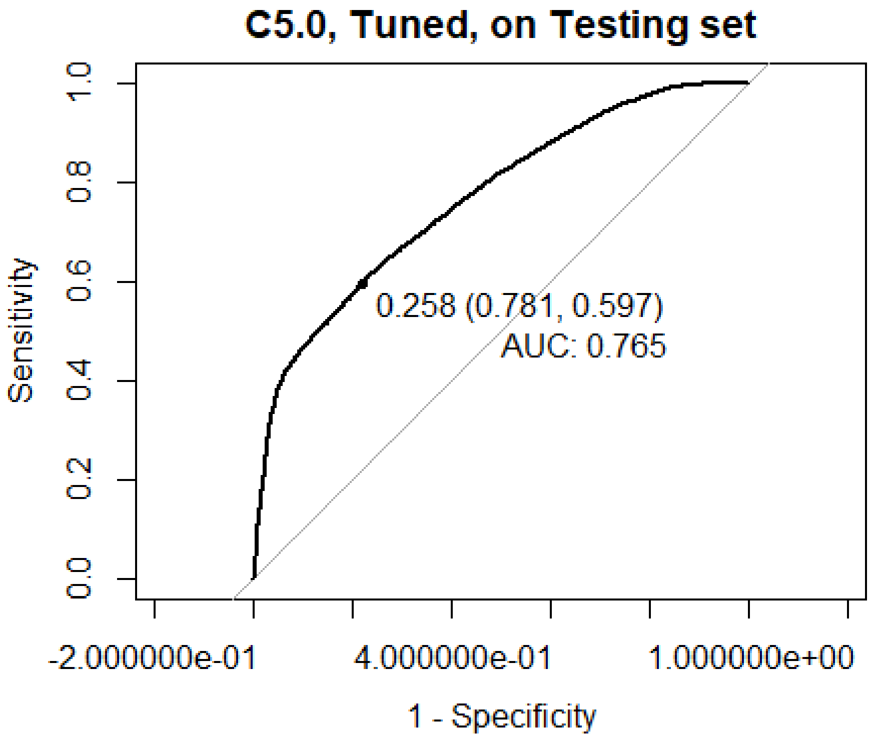
Model	Precision	Sensitivity
C5.0	0.827	0.154

gbm	0.821	0.147
mlp	0.810	0.144

At high levels of precision, C5.0 had the best trade-off between precision and sensitivity. As a result, the final predictions will be made using the tuned C5.0 model on the testing set.

6. Final predictions on test set using best performing model

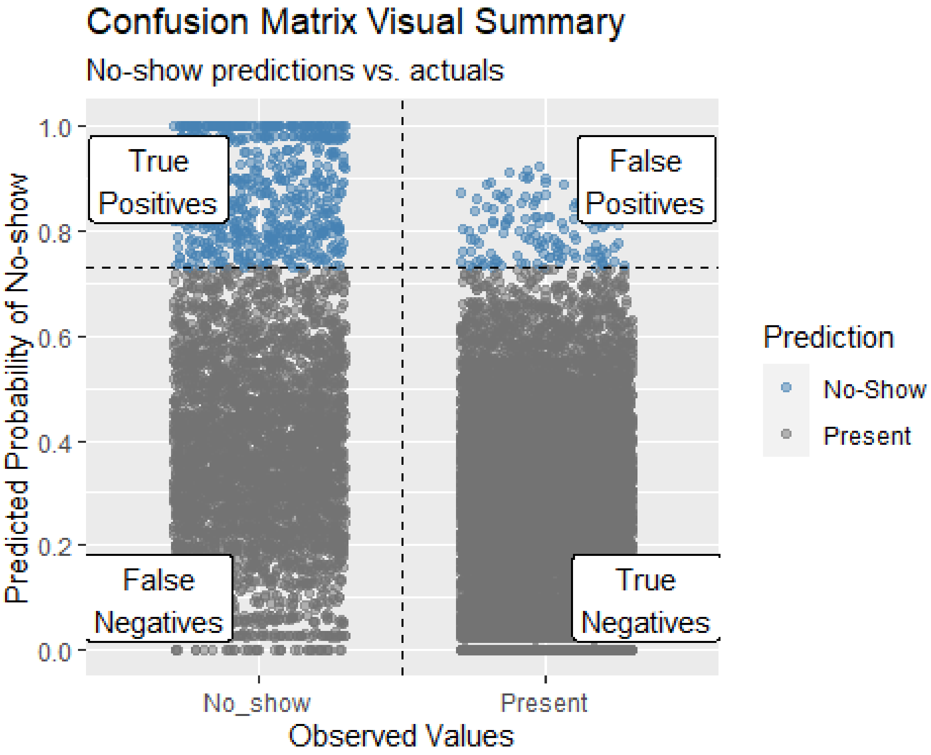
Producing the final model’s ROC curve we observe that the final ROC of 0.765 on the test set is consistent although slightly lower than what was achieved using cross-validated training data.



The final confusion matrix shows that C5.0 correctly predicted roughly 14% of No-shows, at 82% certainty. The higher certainty and lower number of predictions is by design, as mentioned previously.

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction No_show Present
##    No_show     618     135
##    Present    3846   17505
## 
##                 Accuracy : 0.8199
##                   95% CI : (0.8148, 0.8249)
##      No Information Rate : 0.798
##      P-Value [Acc > NIR] : < 2.2e-16
##
##                    Kappa : 0.1897
##
##   Mcnemar's Test P-Value : < 2.2e-16
##
##                Precision : 0.82072
##                   Recall : 0.13844
##                       F1 : 0.23692
##               Prevalence : 0.20195
##           Detection Rate : 0.02796
##     Detection Prevalence : 0.03407
##        Balanced Accuracy : 0.56539
##
##         'Positive' Class : No_show
##
```

The final figure is a visual representation of the confusion matrix, with the probability threshold indicated by the stapled line. It shows the trade-off between Sensitivity and Precision, in that a lower probability threshold would yield more true positives, but also false positives.



Conclusions and Further Considerations

This report completed a best effort analysis to predcict medical appointment No-shows. It did so testing a range of different models, with emphasis on a basic tree model (Rpart), a multilayer perceptron neural net (mlp), and two boosted tree models; Gradient Boosted Model (gmb) and C5.0.

The latter three performed best among all the models tested, with C5.0 achieving the best results. The final model predicted 14% of no-shows at 82% precision. A model such as this can improve the efficiency of a doctor's office.

In the context of this dataset covering about a month, using these predictions could have saved the clinic roughly two thirds of a doctor FTE, assuming 160 hours in a work month. This was estimated by dividing 618 correct predictions by 6 (assuming 6 appointments per hour), resulting in 103 hours saved.

Ethical and practical considerations

Care should be taken when using a model such as this to predict No-shows, as it could result in unintended consequences such as discrimination against groups whose circumstances make them more likely to miss their appointments more often. No patients should face situations where their appointment was cancelled because a model expected the patient not to show up.

Consequently, the way results are used should be developed with care. One suggestion could be to adjust the expected length of appointments in the booking system by their likelihood of being missed, and fill up the total booking time of the course of the day. This would take advantage of the overall predictions for the whole day, instead of cancelling appointments for individuals. Additional work should be done around designing policies for applying machine learning models to improve doctor's offices' efficiency.

Features and further exploration

The model indicated that patient age, past no-show history, time from booking to appointment, and booking frequency are important features for no-show prediction.

The presence of negative values for some variables (i. e. time from booking to appointment) meant that they could not undergo log transformation, despite demonstrating a distribution that would lend itself well to said transformation. Access to data without such negative values could potentially improve results.

Another iteration could split data according to time, so that the model is trained on appointments happening at the beginning of the timeframe, and the testing is done on appointments at the end of the timeframe. This would avoid the issue of having all patients beginning at 0 past no-shows. However, new patients will always have 0 no-shows. The implication from this is that predictions on new patients without any history is less accurate than predictions for individuals with many past appointments.

Finally, the data reflected patients and appointments in Brazil. Results may not be immediately transferable to other contexts.