Jeevan Bandreddi

No-Show Classification Model

1 Introduction

1.1 Motivation

No-shows, when patients miss scheduled appointments, contribute complications to medical organizations, impacting revenue and clinic efficiency. Having constructed no-show rate reports, building a predictive model that indicates whether a patient will no-show gives further insight to clinic staff, knowing which appointments can be double booked or require additional interventions to ensure patient attendance.

1.2 Source of the Data

Dataset contributing to construct the no-show predictive model was obtained from Kaggle. It contains 15 variables and 300,000 medical appointments from 2014-2015 throughout Brazil.

https://www.kaggle.com/joniarroba/noshowappointments/version/1

1.3 Details of the Dataset

Tables below provide details on all possible relevant variables. Variables *AppointmentRegistration* and *ApointmentData* were preliminarily removed from consideration because exact dates serve as poor predictors without transformation and the variable *AwaitingTime* already extracts information from these two fields being the date difference from appointment made date to the scheduled date.

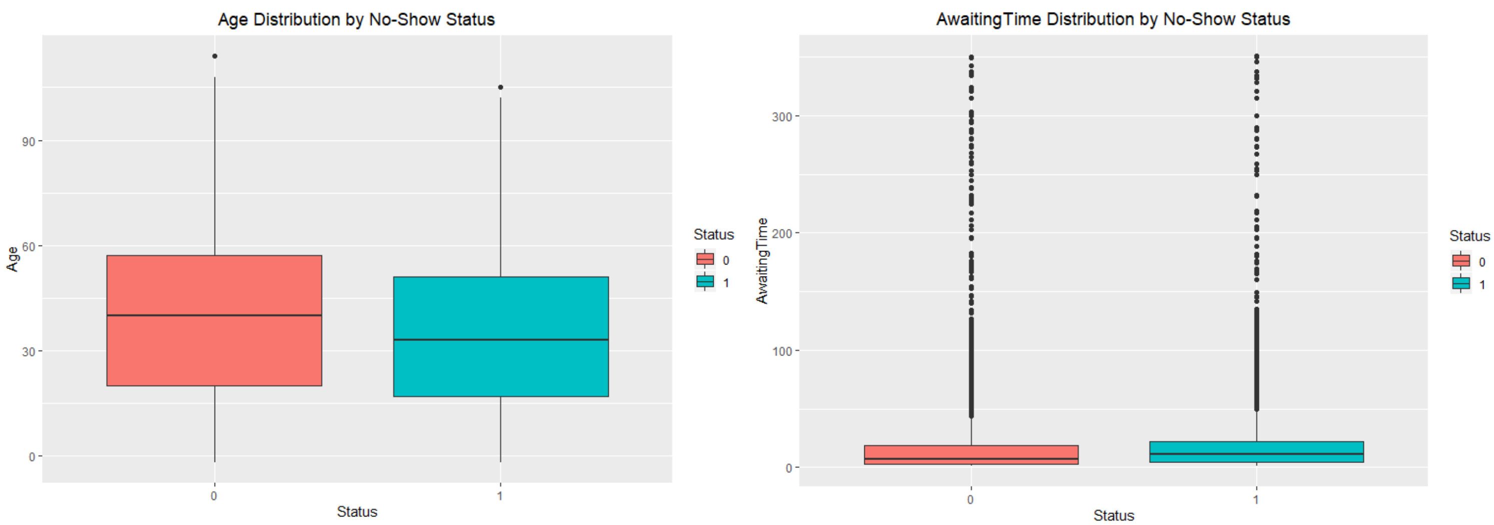
Table 1. Variable Information

Variable	**Description**	**Type**	**Role**
Age	patient age in years	Continuous	Predictor
Gender	gender of patient	Categorical	Predictor
DayoftheWeek	day of week appointment occurred	Categorical	Predictor
AwaitingTime	days difference between appointment made date and appointment date	Continuous	Predictor
Diabetes	indicator - patient’s diabetes status	Categorical	Predictor
Alcoolism	indicator - determines if patient has alcoholism	Categorical	Predictor
HiperTension	indicator - patient’s hypertension status	Categorical	Predictor
Handcap	determines patient’s handicap level	Categorical	Predictor
Smokes	indicator - patient’s smoking status	Categorical	Predictor
Scholarship	indicator - determines if patient has Bolsa Famila (welfare program for the poor	Categorical	Predictor
Tuberculosis	indicator - patient’s tuberculosis status	Categorical	Predictor
SMS_received	determines if text message reminder was sent to patient	Categorical	Predictor
Status	indicator – determines if patient no-showed the appointment	Categorical	Outcome

Table 2. Variable Summary Statistics

Variable	**Summary Statistics**	**No-show rate comparison** *For categorical variables, displays no-show rate for each level. For continuous variables, shows distribution of variable based on Status*
Age	Mean: 37.78 Median: 38 Low: -2 High: 114	*See figure below.*
Gender	F: 201003 *(0.67)* M: 98997 *(0.33)*	F: *0.299* M: *0.309*
DayoftheWeek	Monday: 59777 *(0.199)* Tuesday: 63170 *(0.211)* Wednesday: 63231 *(0.211)* Thursday: 59804 *(0.199)* Friday: 52676 *(0.176)* Saturday: 1338 *(0.004)* Sunday: 4 *(0.000)*	Monday: *0.323* Tuesday: *0.290* Wednesday: *0.297* Thursday: *0.294* Friday: *0.308* Saturday: *0.390* Sunday: *0.250*
AwaitingTime	Mean: 13.86 Median: 8 Low: 1 High: 351	*See figure below.*
Diabetes	0: 276911 *(0.923)* 1: 23089 *(0.077)*	0: *0.307* 1: *0.250*
Alcoholism	0: 292552 *(0.975)* 1: 7448 *(0.025)*	0: *0.301* 1: *0.365*
HiperTension	0: 235237 *(0.784)* 1: 64673 *(0.216)*	0: *0.317* 1: *0.249*
Handcap	0: 294312 *(0.981)* 1: 5183 *(0.017)* 2: 452 *(0.002)* 3: 35 *(0.000)* 4: 18 *(0.000)*	0: *0.303* 1: *0.262* 2: *0.265* 3: *0.257* 4: *0.111*
Smokes	0: 284404 *(0.948)* 1: 15596 *(0.052)*	0: *0.300* 1: *0.351*
Scholarship	0: 270641 *(0.902)* 1: 29359 *(0.098)*	0: *0.296* 1: *0.361*
Tuberculosis	0: 299860 *(1.000)* 1: 140 *(0.000)*	0: *0.302* 1: *0.400*
SMS_received	0: 127783 *(0.426)* 1: 171455 *(0.572)* 2: 762 *(0.003)*	0: *0.303* 1: *0.302* 2: *0.337*
Status	Show-up: 209132 *(0.698)* No-Show: 90688 *(0.302)*	*N/A*

Figure 1. Continuous Variable Distribution by Status



Preliminary data exploration shows differentiation between no-show and present appointments. No-shows tend higher with younger and scholarship patients. Patients with chronic conditions tend to no-show less, while patients with behavioral conditions, like alcoholism and smoking, no-show more. This analysis provides insight into variable inclusion and importance.

Variance analysis identifies three categorical variables with near zero variance in their classification values. *Tuberculosis,* being the most extreme with only 140 patient appointments showing positive for the condition, will be excluded from the models due to its low value of influence.

Table 3. Variance Analysis

> variancecheck

	freqRatio	percentUnique	zeroVar	nzv
Age	2.178358	3.600048e-02	FALSE	FALSE
Gender	2.030354	6.666756e-04	FALSE	FALSE
AppointmentRegistration	1.125000	9.845431e+01	FALSE	FALSE
ApointmentData	1.010610	1.780024e-01	FALSE	FALSE
DayOfTheWeek	1.000966	2.333364e-03	FALSE	FALSE
Status	2.308038	6.666756e-04	FALSE	FALSE
Diabetes	11.993027	6.666756e-04	FALSE	FALSE
Alcoolism	39.278733	6.666756e-04	FALSE	TRUE
HiperTension	3.638659	6.666756e-04	FALSE	FALSE
Handcap	56.783330	1.666689e-03	FALSE	TRUE
Smokes	18.235445	6.666756e-04	FALSE	FALSE
Scholarship	9.218195	6.666756e-04	FALSE	FALSE
Tuberculosis	2141.828571	6.666756e-04	FALSE	TRUE
Sms_Reminder	1.341736	1.000013e-03	FALSE	FALSE
AwaitingTime	1.112364	7.200096e-02	FALSE	FALSE

The final list of predictors:

Age	Handcap
Gender	Smokes
DayoftheWeek	Scholarship
Diabetes	Sms_Reminder
Alcoolism	AwaitingTime
HiperTension	

2 Analyses

2.1 Method of Analysis

Prior to modeling construction, the following pre-processing steps happened for data preparation:

- Predictor categorical variables converted to factors
- Outcome variable *Status* transformed to a 0/1 indicator with 1 being no-show and 0 representing present. Some models require outcome in a 0/1 format.
- Data set contained no missing data. However, four observations had negative age values and were removed.
- *Handcap* and *Sms_Reminder* were condensed to 0/1 indicators as some classification levels contained low count
- Removal of unneeded variables *AppointmentRegistration, ApointmentData,* and *Tuberculosis*
- Numeric continuous variables *Age* and *AwaitingTime* were scaled for placement into the same unit of measure to allow proper distance calculation and comparison
- Data set split into 80%/20% training and holdout groups respectively. Training set used to build the optimal model for each approach while the holdout set tests predictive performance to find the overall optimal model.
- Down-sampling performed on training set to address class imbalance

Classification modeling approaches that can distinguish a binary outcome (two classes) will be evaluated. These are the modeling techniques:

Logistic Regression
K-Nearest Neighbors (KNN)
Linear Discriminant Analysis (LDA)
Quadratic Discriminant Analysis (QDA)
Random Forest
Boosted Tree Model
Support Vector Machines (SVM)

All approaches will be included. Each technique has a distinct methodology to determine class selection. Logistic regression outcomes the probability an observation belongs to a class. LDA, QDA, and SVM build boundaries in the predictor space to maximize separation between the classes, differentiating in boundary type and procedure used in boundary creation. KNN selects the majority value based on the *k* nearest neighbors determined by a measure of distance. Finally, random forest and boosted tree are tree-based approaches that stratify the predictor space into simpler regions using nodes until a class decision occurs.

While some approaches, such as LDA, have assumptions, they are incorporated into model performance. With the premise that models perform poorly when assumptions are not met, this gives no reason to exclude any methods and allows performance to drive usability of the approaches.

To evaluate model performance, the following metrics are calculated from a confusion matrix between observed and predicted outcome values:

Overall accuracy: Number of appointments predicted correctly / Total appointments
True positive rate: Number predicted as no-show / Total no-show appointments

While other accuracy metrics exist, these two metrics have the greatest importance for no-show modeling. High total accuracy gives stability in predictions, allowing medical staff to act appropriately and confidently based on the predicted appointment status. Additionally, it is more imperative to correctly identify no-show appointments over present appointments. The default expectation is patients will show for the appointment. Clinic staff prepare with that expectation in mind so proper flagging of no-show appointments is needed for staff to change course of action.

2.2 Results and Interpretation

Using k-fold cross validation, eight diverse modeling techniques will be evaluated through finding optimal parameter configurations for each technique and evaluating predictive performance against a holdout set to find the overall best-performing model. The first tested approach, logistic regression, serves as the base-line model for comparison. It fills this role well, having no tuning parameters that can influence finding the best variant of the model while also producing coefficients to determine predictor importance. Additionally, it is one of the most studied approaches, seen as a default in classification modeling.

For model evaluation, both metrics, overall accuracy and true positive rate, equally determine predictive performance. Comparison requires observing models that perform well in both metrics and evaluating tradeoffs in gains and losses between the two metrics to find the best performing model. Table below shows metric performance and ranks the models based on joint performance.

Table 4. Model Performance

Modeling Approach	Range of Parameters	Optimal Parameter Configuration	Overall Accuracy	True Positive Rate	Rank
Logistic Regression	*N/A*	*N/A*	55.83	57.22	3
KNN	**K**: 1, 3, 5	**K**: 5	54.20	53.22	5
LDA	*N/A*	*N/A*	55.79	57.27	4
QDA	*N/A*	*N/A*	61.74	36.50	6
Random Forest	**Number of predictors**: 2, 9, 16	**Number of predictors**: 2	55.63	61.09	1
Boosted Tree Model	**Number of trees**: 50, 100, 150 **Interaction depth**: 1, 2, 3 **Shrinkage**: 0.1	**Number of trees**: 150 **Interaction depth**: 3 **Shrinkage**: 0.1	56.92	58.47	2
SVM – linear	**C**: 0.1, 1, 10, 100	**C**:	*N/A*	*N/A*	*N/A*
SVM – radial	**C**: 0.1, 1, 10, 100 **Gamma**: 0.5, 1, 2, 3	**C**: **Gamma**:	*N/A*	*N/A*	*N/A*

As shown in the results above, there are no metrics for the SVM approaches. After running for two hours with no output and evaluating against time and resource constraints, the decision was made to drop these computationally heavy modeling approaches from consideration.

Predictor importance determines a variable's usefulness in influencing the outcome value. Composite ranking using coefficient magnitude and relative influence metrics from the top three performing models determined the most influential predictors. For factor predictors with multiple values, the max coefficient was taken into consideration when determining within model rank.

Table 5. Predictor Placement

Variable	Logistic Regression	Boosted Tree	Random Forest	Composite Score
Age	4	1	1	6
Gender	9	8	9	26
DayoftheWeek	1	6	10	17
AwaitingTime	6	2	2	10
Diabetes	11	10	7	28
Alcoolism	5	7	8	20
HiperTension	8	9	3	20
Handcap	10	11	11	32
Smokes	2	4	6	12
Scholarship	3	5	4	12
SMS_received	7	3	5	15

Table below shows predicator importance in decreasing order. Ties were broken based on higher ranking in the best performing model, random forest.

Table 6. Predictor Ranking

Rank	Variable	Description
1	Age	patient age in years
2	AwaitingTime	days difference between appointment made date and appointment date
3	Scholarship	indicator - determines if patient has Bolsa Famila (welfare program for the poor
4	Smokes	indicator - patient's smoking status
5	SMS_received	determines if text message reminder was sent to patient
6	DayoftheWeek	day of week appointment occurred
7	HiperTension	indicator - patient's hypertension status
8	Alcoolism	indicator - determines if patient has alcoholism
9	Gender	gender of patient
10	Diabetes	indicator - patient's diabetes status
11	Handcap	determines patient's handicap level

Principal component analysis lacks feasibility within a data set containing only 2 continuous variables out of 11. Using only two predictors loses out on large amounts of valuable information disabling the ability to create efficient components for analysis. Figure below shows components generated from only using *Age* and *AwaitingTime*.

Figure 2. Principal Component Analysis

```
Importance of components:
                          PC1    PC2
Standard deviation     1.0099 0.9892
Proportion of Variance 0.5104 0.4896
Cumulative Proportion  0.5104 1.0000
```

2.3 Conclusions

Random forest produced the strongest performing model. It correctly predicted no-show appointments best with a true positive rate at 61.09%, beating out the second-best model by 2.62%. While QDA had the best overall accuracy, its poor aptitude identifying no-show appointments removed it from contention. Random forest maintained close range for overall accuracy with the other models while pulling ahead with its true positive rate.

However, while random forest was the best-performing model within the set, its accuracy rate fairs poorly in the absolute sense, preventing the model from confidently be given to clinic management decision makers to use. The high amount of errors would cause workflow issues as clinic staff make decisions, such as double booking a perceived no-show appointment, based on false information. Model improvement needs to occur internally prior to live implementation. This includes evaluating other modeling approaches, gathering other potentially more influential variables outside this data set, or making changes to the current model such as removing low impact variables, transforming variables, or further optimizing parameters.

While a highly successful model did not generate from this analysis, it identified key predictors influential to no-show prediction that decision makers can review. The top five predictors are *Age, AwaitingTime, Scholarship*, *Smokes,* and *SMS_received.* Younger, scholarship and smoking patients are the best focus groups regarding efficient attempts to reduce prevalence of no-shows. Additionally, longer wait times from appointment creation to being seen and appointments without SMS reminders are more likely to no-show. Clinics can focus on shortening the days difference and making sure all appointments get reminders.

3 Lesson Learned

Lessons learned focus on resource constraint evaluation and viewing modeling as an iterative process. While building the models, I incurred multiple issues with model processing taking too long. In an ideal world, this provides no concern, but in the real world where time is a valuable resource, I juxtaposed running the most exhaustive search to gather optimal results versus making efficiency choices for good results. As Voltaire said, “Don’t let the perfect be the enemy of the good”. To achieve good, completed results, I cut down the range of parameters tested and decided to remove models that were computationally taking too long.

I also learned acceptance of initial poor results as modeling requires trial and error. My no-show models had poor accuracy numbers but still important information was learned about no-shows and the models that can be used for improvement in future iterations. Modeling is a cyclical learning process.