Predicting Medical Appointment No-show status

Andrew Barber

University of Michigan

SI 618

Fall 2018

Predicting Medical Appointment No-show status

Motivation

Healthcare is expensive; on average, an hour of a physician's time will cost $200.[1] This may come as no surprise, but the cost of that time slot remains the same regardless of if there is a patient filling it. In 2017, missed appointments cost the United States healthcare system a staggering $150 billion.[1] To help mitigate these costs, it is important to know well ahead of time whether the patient will be attending their appointment. Without explicit confirmation from the patient, this can be difficult to anticipate, however, the aim of this project is to understand and predict patient no-shows in an attempt to inform clinicians so they can make adjustments to their schedule, ultimately reducing overall healthcare costs.

To accomplish this goal, we will take four basic questions regarding the dataset:

1) What is the relationship between "No-show" status and the given continuous variables: age and appointment wait time (time between "Appointment Day" and "Scheduled Day")? What are the distributions of these variables, are there any outliers?
2) What is the relationship between "No-show" status and the given categorical variables: "Gender", "Neighborhood", "Scholarship", "Hypertension", "Diabetes", "Alcoholism", "Handicap" and "SMS received"?
3) What are the principal components of the dataset. How would they be described?
4) Given the above information, can you predict the likelihood of a patient being a no-show? How accurate is the model?

Dataset

The dataset used for this project was adapted from the "Medical Appointment No Shows" dataset on Kaggle: https://www.kaggle.com/joniarroba/noshowappointments. It includes a file in csv format containing a total of 15 columns and 110,527 rows – each row representing a scheduled appointment. The only continuous variable originally contained in the dataset is age, the rest are categorical variables. Important categorical variables to note include gender (classified binarily), "ScheduleDay" (or day on which the patient scheduled their appointment), "AppointmentDay" (or day the patient scheduled their appointment for) and "Neighbourhood" (the neighborhood the patient is from). The rest of the notable variables include medical status information such as hypertension status, diabetes status and alcoholism. The dataset includes appointments scheduled for dates from April 29, 2016 through June 8, 2016.

Methods

Data Manipulation

To begin with, I loaded the csv file into a dataframe. I then corrected any typos in the column headers (e.g. "Hipertension" to "Hypertension") and changed the "ScheduledDay" and "AppointmentDay" types from string objects to datetime objects. This allowed me to create custom variables based on the day of the week appointments were scheduled/scheduled for as well as the hour of the day the appointment was scheduled. Unfortunately, the data was not so granular as to specify the hour of the day the appointment was scheduled for. When necessary, I also created numerical representations of data such as for "Neighbourhood". Besides the missing time-of-day information from the "AppointmentDay" column, there was no missing or

incomplete data to contend with. All additional data manipulation was limited to filtering out columns for principal component analysis and classification.

Workflow of Source Code

To complete my exploratory data analysis (EDA) tasks as outlined in the *Motivation* section, I filtered the data into two dataframes according to variable type – categorical and continuous. The plots I used to analyze the continuous variables included a distribution plot, QQ plot, lag plot and regression plot. As I am interested in comparing these variables to the binary "no-show" variable, I also conducted an ANOVA for my categorical variables to discern any statistically significant relationships between them and the "no-show" status of the patient. For my categorical variables, I conducted a χ^2 analysis and as with the continuous variables, comparing them to "no-show" status. To visualize these results, I created contingency table heatmaps of observed and expected counts for each of the variables of interest.

For the dimensionality reduction and principal component analysis (PCA) portion of the analysis, I filtered the data into 13 columns of interest including extracted features I created earlier. I then performed PCA on this data and analyzed the results using explained variance values and a scree plot to determine the number of components contained in the dataset. I then created a principal component by feature table to determine the influence of each feature on each principal component.

To complete the classification portion of the project I analyzed results based on my 13 features of interest, the same features mentioned in the PCA section. These features utilized all columns in the original dataset in some capacity excluding "PatientId" and "AppointmentId". In addition to these features, I also wanted to evaluate the results of classification using principal components as features. I hoped to evaluate both the performance of the principal component

features compared to the un-reduced data as well as the speed of execution to consider potential viability in real-world applications. I looked at two classifiers for this project: a random forest model and a support vector machine model (sklearn's SGD classifier). I tuned each of these models using grid search with cross validations of 5 and 10.

Challenges:

The EDA portion of my analysis did not present many obstacles. Besides the creation of new features in the dataset, the task was relatively straightforward. Visualizing my PCA results did prove challenging. I hoped to visualize my results as an aggregated data perceptual map. However, the scale required to include all data rendered the visualization useless. Because of this, I was forced to settle for a standard biplot, representing feature vectors projected onto a principal component 1 by principal component 2 graph. All data was also represented in this biplot behind the vectors.

The largest challenge encountered involved the results obtained from my classifiers. In the original data, the percentage of "no-shows" was only 20.2% compared to 79.8% of patients who showed up to their appointment. This means that any classifier predicting that patient will always show up immediately obtained 79.8% accuracy. This is exactly what happened with my models and limited them in predictive power. I therefore decided to re-sample the data so that there was a proportionate number of "no-shows" and "shows" in the data. This limits the utility of the model for real-world application but establishes a better framework for future directions regarding this classification task.

Analysis and Results

Question 1: EDA of Continuous Variables

As mentioned above, I analyzed four different plots for my continuous variables: distribution plots, QQ plots, lag plots and regression plots. I analyzed my two continuous variables, wait time (or number of days between when the patient scheduled appointment and for which day they scheduled it) and age based on 'no-shows" and "shows" for these categories. On the whole, I found that the wait times were heavily right-skewed for both groups with no linear relationship between the length of the data and the distribution of the data. A sample of these results can be seen in *Figure 1* below. For age, the distribution was largely uniform, although a large number of patients were between 0 and 1 year of age.

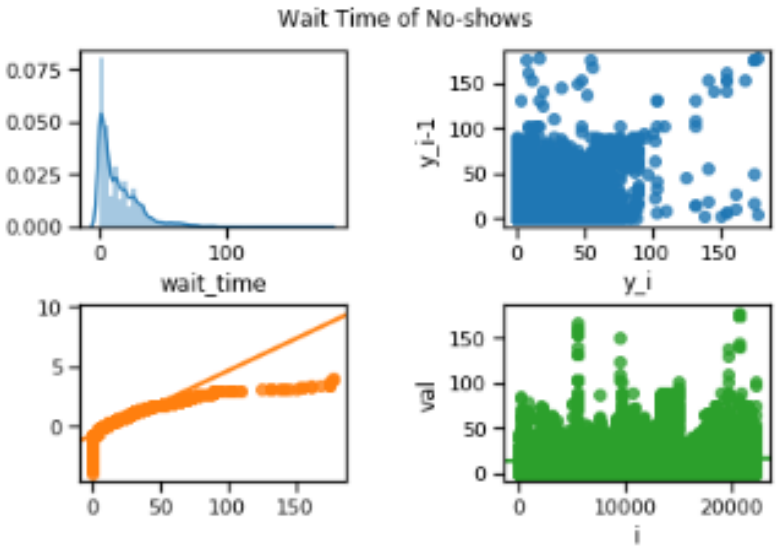

Figure 1: Sample of exploratory visualizations – wait time for "no-shows"

ANOVA results revealed statistically significant differences in "no-show" and "show" proportion based both on age and wait time. The inferences we can make based on the

corresponding regression results are two-fold: the older you are, the more likely you are to go to your appointment and the farther away you schedule your appointment, the more likely you are to miss your appointment. A sample of these regression results can be seen in *Figure 2* below.

OLS Regression Results

Dep. Variable:	Age	**R-squared:**	0.004
Model:	OLS	**Adj. R-squared:**	0.004
Method:	Least Squares	**F-statistic:**	403.6
Date:	Wed, 12 Dec 2018	**Prob (F-statistic):**	1.32e-89
Time:	01:24:34	**Log-Likelihood:**	-5.0371e+05
No. Observations:	110527	**AIC:**	1.007e+06
Df Residuals:	110525	**BIC:**	1.007e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	37.7901	0.078	486.539	0.000	37.638	37.942
no_show[T.Yes]	-3.4724	0.173	-20.090	0.000	-3.811	-3.134

Omnibus:	18548.621	**Durbin-Watson:**	1.165
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4284.977
Skew:	0.114	**Prob(JB):**	0.00
Kurtosis:	2.063	**Cond. No.**	2.61

Figure 2: Ordinary least squares regression model comparing age and "no-show" status

Question 2: EDA of Categorical Variables

Exploratory data analysis for categorical variables consisted of χ^2 tests for each variable compared to "no-show" status as well as contingency table heatmaps for both observed and expected values as a means of visualizing the results. A sample of this output can be seen in *Figure 3* below. A brief description of the variables explored in this section include: gender, receipt of welfare scholarship, hypertension status, diabetes status, alcoholism status, degree of

disability (on scale from 0 to 4), if the patient received one or more messages reminding them of their appointment, neighborhood in which the patient lives, day of the week from Monday to Saturday the patient scheduled their appointment, day of the week the patient scheduled their appointment for and hour of the day at which the patient scheduled their appointment. Of these, all features returned significant differences between observed and expected values except gender, alcoholism, disability level and day of the week on which the patient scheduled the appointment.

What I find to be most interesting of these findings is that the distribution of short message service (SMS) reminders to patients regarding their appointment returned an F-statistic of 1766 making it the most significant result of any of the variables. The surprising thing is that the results suggest that patients who receive an SMS are significantly less likely to show up to their appointment. I am not sure how to explain this finding. Other interesting findings include that hypertension and diabetes status also negatively predict appointment attendance.

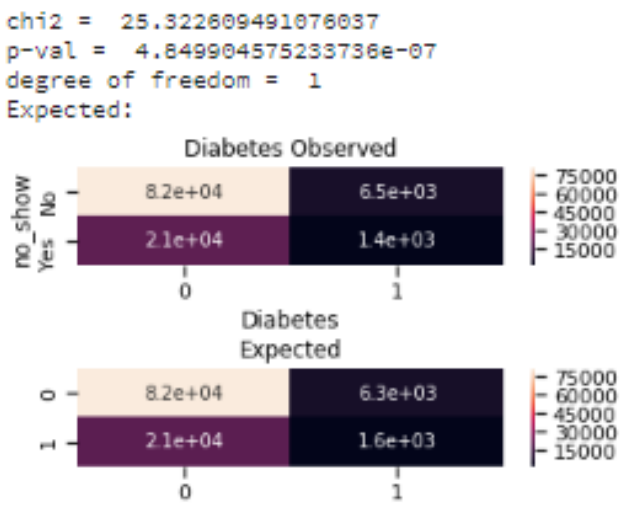

Figure 3: χ^2 test and contingency table visualization for diabetic patients by "no-show" status

Question 3: Principal Component Analysis

For this section, I wanted to look at the principal components of the dataset. I did a PCA analysis on the continuous and categorical variables I analyzed above and then determined the number of components using a scree plot of the explained variance. Although six variables had variance over one, I concluded based on the scree plot that there are four principal components in the dataset. I then plotted my results in a PCA biplot (*Figure 4* below).

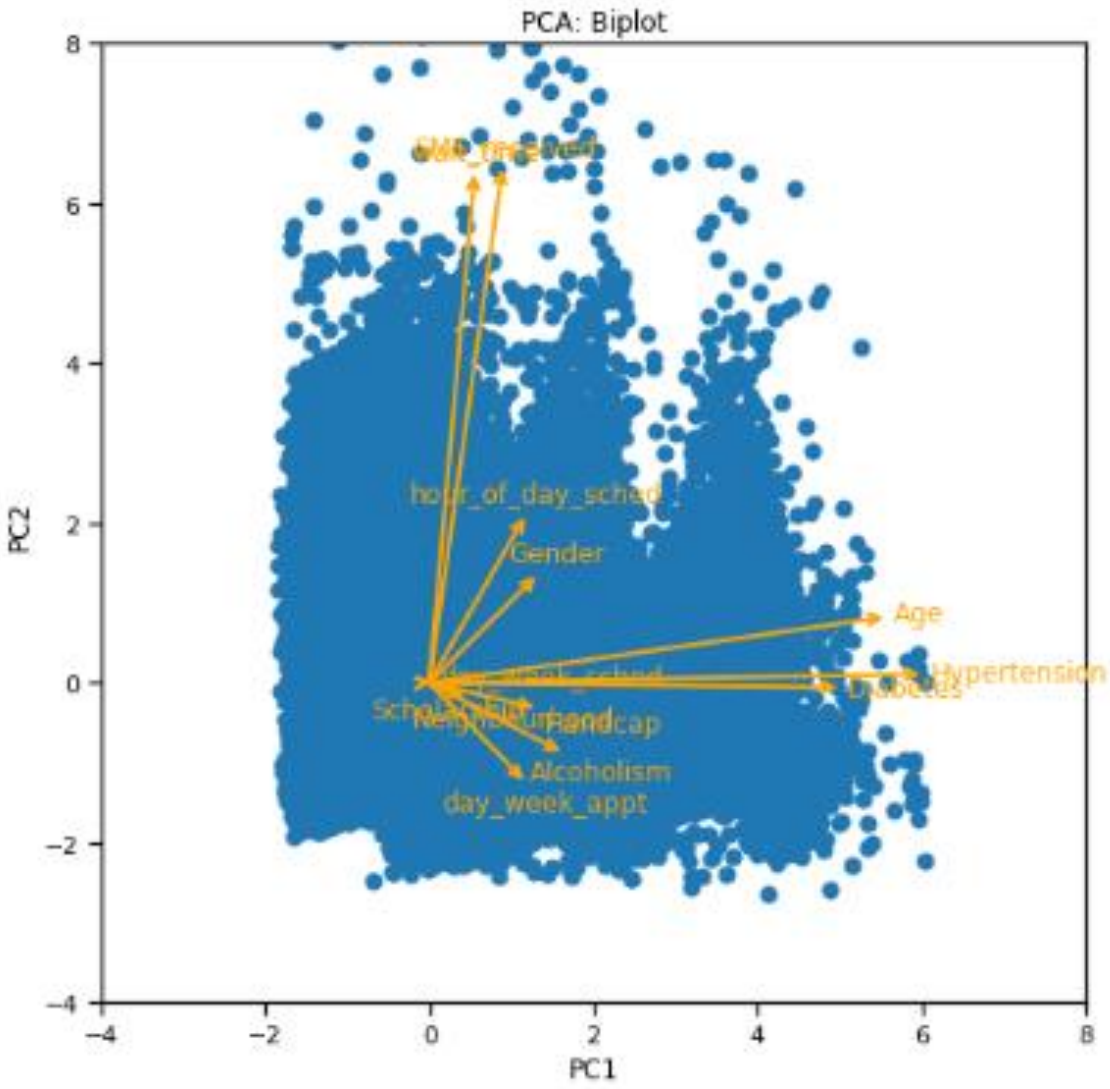

Figure 4: PCA biplot

From the component results I came to the following conclusions regarding my four components: component 1 represents features related to poor health status such as age and hypertension, component 2 represents how likely the patient is to recall the appointment based on factors such as wait time and reception of reminder message(s), component 3 represents the day of the week the appointment and the scheduling took place, and component 4 represents the remaining components. Because of the lack of readability of the biplot, I have also included more details on the composition of the four principal components in *figure 5* below.

	PC1	PC2	PC3	PC4
Gender	0.097035	0.151347	-0.005438	-0.670871
Scholarship	-0.071209	-0.044485	-0.006843	-0.576248
Hypertension	0.609893	0.003448	-0.012104	-0.034213
Diabetes	0.508360	-0.016224	-0.015388	-0.042402
Neighbourhood	-0.020382	-0.053581	0.079091	-0.198342
Alcoholism	0.121201	-0.121533	0.000609	0.309542
Handcap	0.139184	-0.060456	-0.008760	0.125913
SMS_received	-0.018636	0.663302	0.093183	0.008519
Age	0.564720	0.078572	0.000917	0.025268
wait_time	-0.021390	0.656498	0.101586	0.076473
day_week_appt	0.015741	-0.159229	0.680282	0.011793
day_week_sched	0.012961	-0.002200	0.712215	0.009501
hour_of_day_sched	-0.022524	0.225947	-0.064576	0.238963

Figure 5: Results of principal component analysis

Question 4: Classification of "No-shows"

As outlined in the *Methods* section, I performed classification of "no-show" status using two models (random-forest and stochastic gradient descent (SGD)) and two feature inclusion methods (using 13 continuous/categorical variables analyzed above and using principal components model from PCA). I performed a grid search using cross validations of 5 and 10 for each method in order to tune the parameters and have displayed the highest accuracies for each method in *Table 1* below. The main finding for this analysis was that the random-forest classifier outperformed the SGD classifier and using principal components at least marginally increased the accuracy of both models.

The main limitation surrounding these results involves the method of resampling the data. As mentioned in the *Methods* section, the distribution of "no-shows" to "shows" was a disproportionate 20.2% to 79.8%. This made the training of the models difficult as they would always classify a patient as "show", therefore never surpassing the baseline of 79.8% accuracy. I therefore used an under-sampling method to even out the distribution of "shows" and "no-shows" in the dataset.[2] This allowed for more discriminative classifiers but at the expense of real-world application. The accuracies obtained do not come close to a baseline 79.8% accuracy from the original data but they do handily beat the baseline based on the assumption of equal chance that a patient will show up to their appointment or not. This definitely does not apply to the real world, but the strengths in the results below fall under the methods used to obtain them as well as a basis for further analysis and optimization; the increase in accuracy based on PCA is promising and I believe this will only become more pronounced with further feature extraction and external sampling.

Another useful finding comes from the speed of computation for the different models. Although I would not define the differences in results between PCA and 13-feature-based methods conclusive, it is safe to say that after training the model, the output of predictions for the SGD classifier are much faster than for that of the random-forest model. The trade-off there would certainly be the accuracy of the model, but it is useful information to consider upon the possible implementation of real-world applications.

Table 1

Classification Results

Model	Highest Accuracy	Accuracy using 13 features	Speed of 13 features computation (milliseconds)	Accuracy using principal components	Speed using principal components (milliseconds)
Baseline	50%	_-	-	-	-
Random-Forest	67.46%	67.38%	~127	67.46%	~160
SGD	61.28%	58.01%	~7.84	61.28%	~2

References

1) Gier, J. (n.d.). Missed appointments cost the U.S. healthcare system $150B each year. Retrieved from https://www.healthmgttech.com/missed-appointments-cost-u.s.healthcare-system-150b-year

2) Alencar, R. (n.d.). Resampling strategies for imbalanced datasets. Retrieved from https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

3) JoniHoppen. (2017, August 20). Medical Appointment No Shows. Retrieved from https://www.kaggle.com/joniarroba/noshowappointments/home