# Predicting Patient Appointment No-Shows

Catherine Magsino

Patient no-shows and last-minute cancellations come at a very high cost. On average, health care practices experience about a 19% average no-show rate. With the average cost of an unused appointment slot costing about $200, no-shows total to about $150 billion per year for the US healthcare system. More importantly, missed appointments have a high negative impact on the hospital staff and patients. The time wasted on an unused appointment could have been spent treating another patient and improving quality of care. The goal of this project was to create a model that predicted whether or not a patient would show up to an appointment, and to provide a recommended course of action to reduce the amount of missed appointments.

## Data

The data I used for the project was retrieved from a Kaggle dataset. This dataset consists of 111K appointments for 62K distinct patients across 81 different neighborhoods. The time period spanned from April to June 2016.

While the target of the model was to classify a patient as a no-show/show, my final dataset consisted of 13 features, as the original dataset required a bit of feature engineering. This included:

- Mapping the No-Show field to 0/1
- Calculating the number of days between the appointment schedule date and actual appointment date
- Calculating the cumulative number of appointments each patient previously had prior to the appointment
- Calculating the cumulative number of missed appointments each patient previously had prior to the appointment
- Changing the handicap field to whether or not the patient has a handicap
- Removing outliers

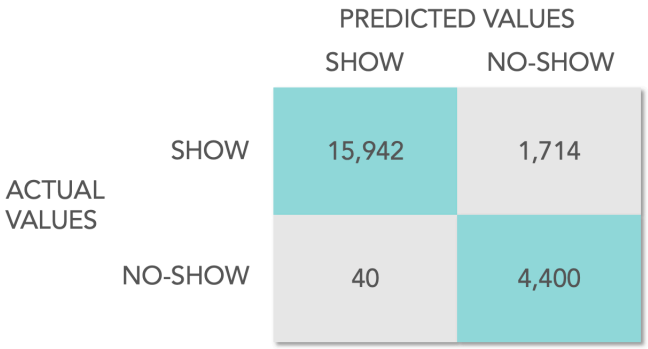
In addition, the features were normalized using StandardScaler. Since the data consisted of an 80/20 split between no-shows and shows, I balanced the data using random oversampling.

## Model Selection

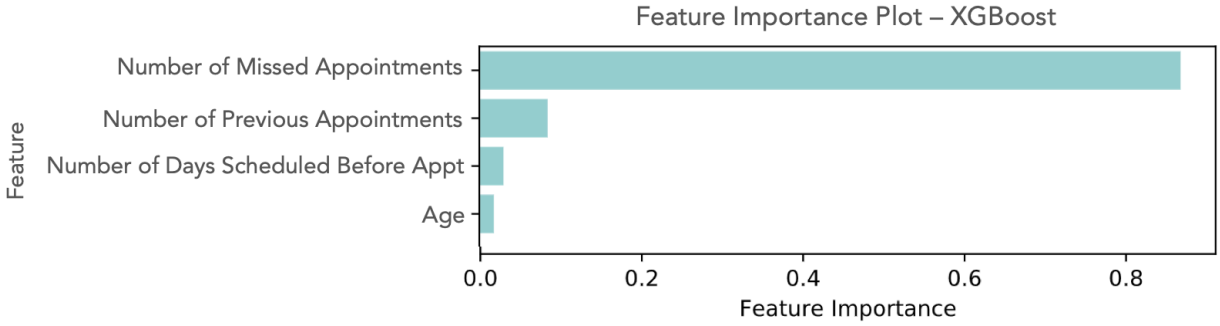
As I worked on choosing the best model for predicting patient no-shows, I ran the data through 7 different statistical models and compared the results using the balanced accuracy metric. The XGBoost model gave the highest balanced accuracy of 0.95, while Support Vector Machines came very close at 0.94.

| Model | Balanced Accuracy |
|---|---|
| Logistic Regression | 0.937 |
| Random Forest | 0.920 |
| Decision Tree | 0.886 |
| XGBoost | 0.945 |
| K Nearest Neighbors | 0.864 |
| Support Vector Machine | 0.941 |
| SVM GridSearchCV | 0.856 |

Since balanced accuracy is calculated by averaging the recall and specificity, this metric was based on my model's ability to predict 99% of no-shows and 90% of patients that showed up to their appointments. I preferred to maximize my recall in order to capture almost all of the patient no-shows.




Looking at feature importance, the most importance feature across all models was the number of appointments missed in the past. This shows that past behavior is a big indicator of whether or not someone will show up to their appointment. Other important features included the number of previous appointments, the number of days between scheduling an appointment and the actual appointment, and the patient's age.



## Recommendation

Based on my results, I would suggest the following recommendations:

- Fee Policy – Since the number of missed appointments was the most important feature of the model, I would recommend instilling a fee policy. This will help with reducing the number of repeat no-shows.
- Automated reminder system with two-way communication – In order to minimize the amount of manual labor required from the staff, I'd suggest creating an automated reminder system that allows the patient to confirm, cancel, re-schedule, or ask any questions.
- Extra reminders for those predicted to not show up – Using the model, I would recommend sending extra reminders to predicted "no-show" patients in order to give them additional opportunities to confirm their appointment attendance.
- Waitlist – Have a waitlist as back-up to fill appointment slots that open up at the last minute.

## Future Work

In order to improve my model, I would look at additional data including neighborhood demographics, accessibility to public transportation, and a longer time period. I would also look into modeling the most effective form of reminder communication.