Predicting No-show Medical Appointments

Sanya Bathla Taneja
University of Pittsburgh

April 2019

Abstract

No-show medical appointments include patients who make appointments at a medical or health care facility and neither use nor cancel the appointment. This project uses the no-show medical appointment dataset obtained from Kaggle.com to predict no-shows at a healthcare facility. Among the variables in the dataset, Age, WaitDays and SMS received are seen to have the highest impact on the no-show rate. Logistic Regression, Random Forest and AdaBoost models are used for prediction with a train and test set. Experiments are also performed with a subset of important features derived from random forest as well as modified thresholds using the ROC curve. The results obtained as well as insights from importance of features can be used for interventions to reduce the no-show rates. Although the results are promising, there is room for improvement in prediction using more data and considering other features that may affect the no-show rate.

1 Introduction

No-show medical appointments include patients who make appointments at a medical or health care facility and neither use nor cancel the appointment. With the current explosion of technology, it is very easy to schedule appointments with just a few clicks. However, studies show that people do not necessarily adhere to these appointments due to a variety of reasons. The no-show rate may vary across different settings but an average of 42 percent appointments become no-shows (Macharia, Leon, Rowe, Stephenson, & Haynes, 1992). Given the large number of appointments per day in primary care hospitals, this is an astounding number.

Time and money are critical resources in every hospital and health care facility. Due to patient no-shows, doctors lose their valuable time and it is downright costly for the health care practice. One study estimates the health care industry loses about 150 billion dollars to patient appointment no shows each year — about 200 dollars in co-pays, reimbursements, and overhead for every hour-long time slot that goes unfilled (“5 Reasons Patients No-Show for Healthcare Appointments | PatientPop,”). If these patient cases can be predicted in advance, there would be significant saving in resources and thus increase in efficiency of healthcare practices by scheduling more patients, giving time slots to other patients etc. Moreover, if we can identify the features of patients who show up to the appointment, we can introduce interventions to successfully reduce the no-show rate.

This project aims to predict whether a patient will show up for an appointment or not. The dataset used is publicly available on Kaggle.com (https://www.kaggle.com/joniarroba/noshowappointments), released in May 2016, consisting of appointments made by patients along with patient and appointment details. The details of the appointment and patient are used as features to model the prediction as a supervised learning problem in machine learning. No-show is taken as the target variable. The dataset reflects the general no-show rate in literature as the number of no-shows are 30 percent of the total appointments.

The project uses Logistic Regression and Random Forest models for classification and prediction. The results are compared to the majority baseline model. Further, features are weighted by importance using the random forest model and subset of data with only the important features is used for classification and prediction using the Random Forest and AdaBoost models. All models are evaluated on a holdout test set using the accuracy, precision, recall and F1 score. Additionally, the threshold of the models is varied using the ROC curve to optimize precision and recall and compare performance to the earlier models. The highest prediction accuracy is obtained with the Logistic Regression and Random Forest models with default classification threshold (0.5). However, these models give poor precision and recall due to the highly imbalanced nature of the data. The best performance in terms of precision, recall and F1 score is observed in Random Forest and AdaBoost models with modified thresholds. Thus, the model predicts no-shows with reasonable precision and recall scores. There are certain features present in the dataset which do not contribute to the prediction and are seen to have no impact on the no-show rate. Age, waiting time and received SMS are seen to have the highest impact on the no-show rate.

2 Related Work

There have been several qualitative studies to discover the reasons for no-show medical appointments. (Lacy, Paulman, Reuter, & Lovejoy, 2004) and (Barron, 1980) analyze these reasons and find that these may include financial concerns, discouragement due to long wait times for the appointment, logistic issues, or simply human forgetfulness. (Elvira, Ochoa, Gonzalvez, & Mochon, 2017) use data from ancillary appointments and consultations at a hospital in Madrid to predict no-shows using the Gradient Boosting algorithm. Their experiments correspond to the results obtained in this project in terms of prediction scores as well as the relative importance of features in the data, although there is some difference in the patient details available in the different datasets.

Logistic regression is a popular supervised learning approach for classification. It also provides a good model for comparison against the majority baseline. Random Forest classifier is an ensemble method based on the decision tree model which uses averaging to improve the predictive accuracy in classification and regression. Random forest model is also used to acquire the relative importance of features based on the classification model. AdaBoost, short for Adaptive Boosting algorithm, is also an ensemble classifier that uses iterative training to improve the classification of the model using multiple weak classifiers.

3 Methodology

3.1 Data

The dataset consists of 110,527 appointments with the following details for each appointment:

- PatientID
- AppointmentID
- Gender

- Appointment Date
- Scheduling Date
- Age (in years)
- Neighborhood
- Hypertension
- Alcoholism
- Diabetes
- Handicap
- SMS Received
- Scholarship
- No-show

PatientID and AppointmentID are unique identifiers. Gender, hypertension, alcoholism, diabetes, SMS received and no-show are categorical variables while age and handicap contain discrete values. Both appointment date and scheduling date are datetime objects. All features except the PatientID, AppointmentID and Neighborhood are used for prediction in this project.

3.2 Preprocessing and Exploratory Data Analysis

Before classification, data is preprocessed to remove any discrepancies and handle the combination of discrete and categorical variables. There are a few patients with age values less than 0. These are assumed to be errors and the instances are removed from the dataset. To use the appointment date and scheduling date, a new variable called 'WaitDays' is created to reflect the waiting time between the date of scheduling and date of appointment. Again, values with WaitDays less than 0 are removed from the dataset.

Further preprocessing includes discretization of the Age and WaitDays variables. This is required to ensure that all values in the data are on a similar scale in classification model. All categorical variables are converted to one hot encoded variables before classification. The dataset is divided into train and test split with a 25 percent used as a holdout test set. This is done with stratified splitting to allow the distribution of classes in the split to remain the same as in the original data. Training of models is done on the train data and all models are evaluated on the same test set.

The exploratory data analysis includes plotting of variables to show the distribution as well as plots against the target variable. Figure 1 shows the distribution of age in the dataset and Figure 2 shows the values of the WaitDays variable. Figure 3 plots the probability of a patient showing up to the appointment against the other variables i.e. Handicap, SMS received, Hypertension, Diabetes, Alcoholism, and Scholarship.

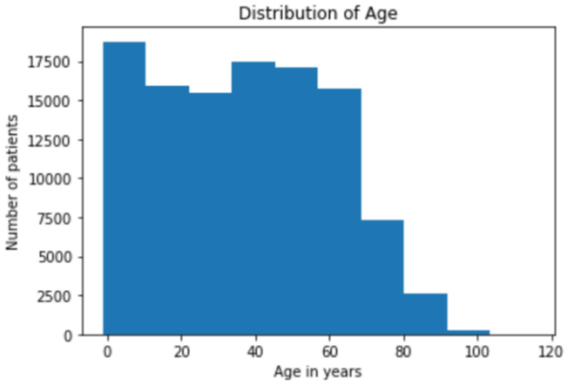

Figure 1: Distribution of age in data

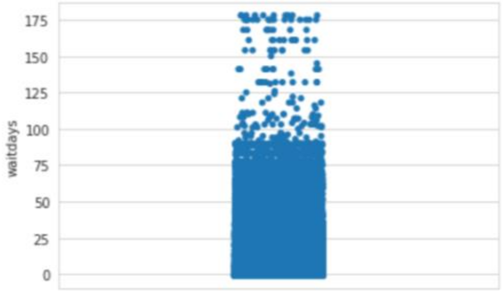

Figure 2: Plot of WaitDays

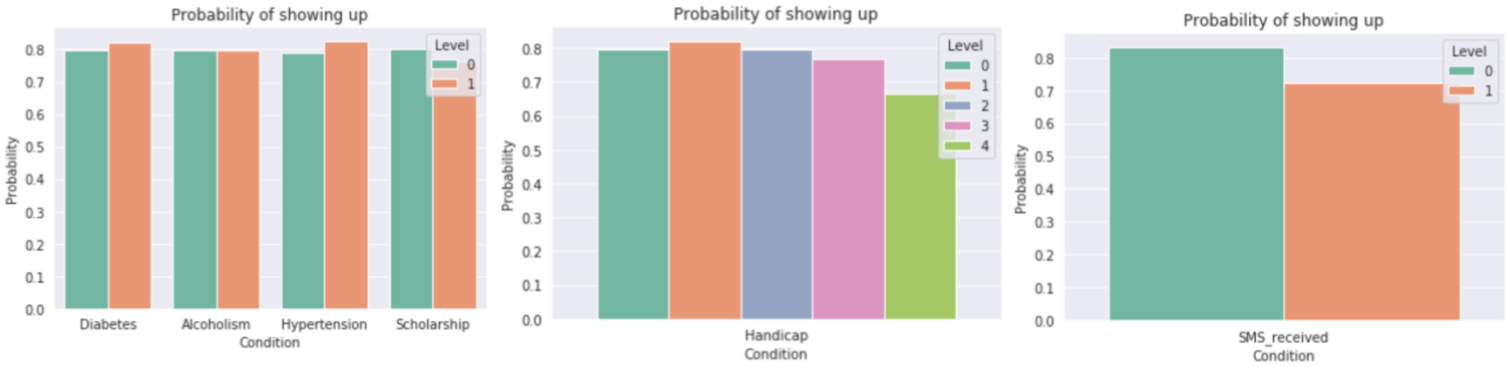

Figure 3: Plots of variables Diabetes, Alcoholism, Hypertension, Scholarship, Handicap, SMS received with the probability of showing up

3.3 Classification

There are four major classification and prediction experiments performed in the project. All models are fit using 5-fold cross validation on the training data. The classification, prediction and evaluation is conducted using the Scikit-learn (Pedregosa et al., 2011) package.

- The entire training data is used for classification using Logistic Regression and Random Forest models. The hyperparameters are chosen by searching over options using GridSearchCV which gives the model with the highest score on the given evaluation metric. Both models are refit using the Area Under ROC curve scores.
- Random Forest model fit above is used to acquire the importance of all features in the classification. The weights obtained show that relative importance of each variable in classification in the training dataset. With the highest scores, the variables Age, SMS received, and WaitDays are used to fit the data again using Logistic Regression and Random Forest and compare the performance with the original models. This reduces the number of features in the model while including all the necessary information for classification.
- Due to the imbalanced nature of the dataset, most classifiers are prone to predict the majority class when fit to the data. This leads to poor performance of the classifier, although with high

accuracy due to the majority class prediction. The ROC curve plots the values of sensitivity and (1-specificity) for different values of threshold in prediction. The default threshold for prediction is 0.5. The ROC curve can be used to obtain the optimal threshold for the classification of the given data and thus improve the prediction of the minority class (no-shows in this case). Thus, the random forest model is refit to predict the no-shows using the optimal threshold (modified) and evaluated.

- Lastly, AdaBoost model is fit to the training data to improve the performance and compared to the models fit before. The AdaBoost model is also used for prediction with a modified threshold (as above).

3.4 Evaluation

The prediction models are evaluated using the same holdout test set. For each model, the following metrics are evaluated:

Accuracy = number of correctly predicted instances divided by the total instances (1)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

All the above metrics are calculated using the confusion matrices. We take macro-average of the precision, recall and F1-score values to reflect the predicted instances for both classes. Although accuracy is one of the evaluation metrics reported, it is not taken as the metric to choose the best model as we are concerned with prediction of the minority class. In this case, precision, recall and F1-score give a better picture of the classification performance than the accuracy.

The threshold for prediction is modified using the ROC curve, which is created using the True Positive Rate and False Positive Rate at various thresholds. The intercept of the curve with the tangent at 45 degrees parallel to the no-discrimination line that is closest to the error-free point (0,1) is used to find the optimal threshold for the given model for further prediction.

4 Results and Discussion

The results for the first classification experiment with all given features is presented in Table 1. The accuracy of all the models is same as the majority baseline model. However, we can see that both the Logistic Regression and Random Forest models give higher precision, even though the recall is the same. Thus, there is improvement over the majority baseline in prediction of no-shows.

The weights obtained for relative importance of features from the Random Forest model are given in Table 2. As shown, Age, SMS received and WaitDays have the highest impact on no-show. The subset of the original train data with only these three features is then used for classification and results are shown in Table 3. These results show that the models do not improve by removing the features with low importance. Although the random forest model performs better than both majority baseline and logistic regression, it does not improve over the original random forest model with all the features.

Table 4 shows the results of evaluation after modifying the threshold using the ROC curves. This is done for the random forest models for both the above cases. Further, AdaBoost model is fit to the data and the result with modified threshold is presented in Table 3 as well. The highest precision and recall are observed in the AdaBoost model which predicts the no-shows on the test set. Both the Random Forest and AdaBoost models with modified thresholds outperform the models used before. This implies that in imbalanced datasets, it is crucial to allow the threshold value to change according to the predicted probabilities to improve the performance.

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
Majority Baseline	0.798	0.39	0.5	0.438
Logistic Regression	0.795	0.58	0.5	0.46
Random Forest	0.798	0.74	0.5	0.45

Table 1: Results of prediction on test set with model using all features

FEATURE	IMPORTANCE
Age	0.13
Wait Days	0.75
Gender	0.01
SMS Received	0.07
Handicap	0.01
Hypertension	0.01
Diabetes	0
Alcoholism	0.01

Table 2: Features with importance obtained from Random Forest model

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression	0.798	0.4	0.5	0.44
Random Forest	0.798	0.4	0.5	0.44

Table 3: Results of prediction on test set with model using only top three important features

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Random Forest	0.553	0.55	0.55	0.55
AdaBoost	0.566	0.62	0.67	0.54

Table 4: Results of prediction on test set with modified thresholds

Based on the results, there are several interventions possible to solve the problem of no-shows. Although the models with reduced features do not perform as well as the original models, there is value in assessing the contribution of the features to the classification. As Age, SMS received and WaitDays have the highest relative importance scores, there can be solutions developed to target patients that do not show up due to extremely long delays between scheduling date and appointment date, as well as opportunities to reduce no-shows by delivering reminders through SMS and other ways.

5 Conclusion and Future Work

This project aims to predict the no-show medical appointments using the publicly available dataset on Kaggle.com. The results of classification and prediction are showcased using a number of models including ensemble classifiers such as Random Forest and AdaBoost. The research also presents the advantage of modifying the threshold for prediction in these models to improve performance. Further, the weights of the features give potential avenues to solving the problem of no-shows. In future work, the author aims to incorporate the neighborhood feature in the classification model. There is also scope for analysis of features to predict which interventions may be successful in making people show up to appointments. Finally, the author would also like to conduct the experiments on a more uniformly distributed target variable and compare results to the ones obtained in this project.

References

5 Reasons Patients No-Show for Healthcare Appointments | PatientPop.

Barron, W. M. (1980). Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care*, *7*(4), 563–574.

Elvira, C., Ochoa, A., Gonzalvez, J. C., & Mochon, F. (2017). Machine-Learning-Based No Show Prediction in Outpatient Visits. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(7), 29.

Lacy, N. L., Paulman, A., Reuter, M. D., & Lovejoy, B. (2004). Why we don’t come: patient perceptions on no-shows. *Annals of Family Medicine*, *2*(6), 541–545.

Macharia, W. M., Leon, G., Rowe, B. H., Stephenson, B. J., & Haynes, R. B. (1992). An overview of interventions to improve compliance with appointment keeping for medical services. *JAMA*, *267*(13), 1813–1817.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.