Predicting Appointment Cancellation and Isolating Contributing Factors (Medumo Group 6)

Jessica Sandler, Heather Johnson, Nicholas Pearce, Jacob Kozol
{sandlerj, heathdj, npearce, jkozol}@bu.edu

1. Project Task

When patients miss their appointments, it leaves empty spaces in hospital schedules that could otherwise be used to serve other patients. The goal of our project is to help hospitals predict which patients will not appear at their appointments so they can use that space to help other patients waiting for appointments. We also want to find which demographics are the most effective at predicting this information. We have decided to focus on the prediction model portion as determining which demographics does not involve machine learning techniques.

2. Related Work

[1] and [2] use, respectively, multinomial and binary logistic regression as the primary methods for solving the classification problem of predicting whether or not a patient will cancel/fail to appear at their appointment. [2] also uses L2-norm regularization to prevent overfitting, and 10-fold cross validation to assess the accuracy of their model.

3. Approach

Taking the factors of our data set as X, we will use binary logistic regression since we are only looking for the binary classification of took place (0) or missed appointment (1) with L2-norm regularization to counteract potential overfitting. We will also use the cross-entropy cost function in conjunction with gradient descent to optimize our solution. Because we are in search of the most accurate model, we have decided to expand our project to include multiple binary classification models. Aside from logistic regression, we will also create a support vector machine, a random forest classifier, *and a binary neural network classifier.*

We have now implemented the logistic regression model, a random forest model, a MLP classification model, and support vector machines with four different kernels. Because it is hard to know the exact shape of our data, we tried multiple kernels within the support vector machine in order to see which kernel does the best job of separating the data. Due to this uncertainty, we are also using models that work best for different data shapes. Logistic regression will work well if there is a clear decision boundary whereas a random forest will perform well on non-linear features. A logistic regression has a lower likelihood of overfitting in comparison with a random forest especially since we can use L2-norm regularization. Both the SVM and random forest can be handle a large feature space than a logistic regression so as we add features they may outperform the logistic regression. Additionally, we are trying to improve our MLP by determining what activation function and regularization term are best for our data.

4. Dataset and Metric

Our dataset has 1996 training examples and 222 test examples. Each example currently includes an index ID, Hospital ID, patient ID, registration date, procedure date, and whether or not the patient appeared at their appointment. We were also given engagement data for each patient, as well as some basic demographic and enrollment information for each patient (gender, age, registration for email and SMS notifications). Before we begin to create our model, we should first compile all of this data so that all of the details are in one table, rather than two or three. However, over time we may need to modify our dataset, as one of our goals is to pinpoint which factors have the highest predicting ability.

The features we have so far include the time between registration and appointment date, the patient's age and gender, whether they are signed up for email and/or SMS notifications, the number of days between when they scheduled the appointment and the appointment itself, the number of completed modules, and the number of messages received. We also plan to include if the appointment date is on a weekday or weekend and what season it falls in. We are working on determining whether or not the patient added a family member to receive notifications.

The metric that we will use to evaluate success will be the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR). In this case, the TPR is defined as the number of correctly predicted cancellations/no shows divided by the total number of cancellations/no shows, and the FPR is defined as the number of successful appointments that have been predicted to be cancellations/no shows divided by the total number of successful appointments. The specific metric we will use is the area under the curve (AUC) of the ROC curve. The AUC is equal to the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the classifier. Our goal is to achieve an AUC between 0.8 and 0.9.

5. Preliminary Results

We currently have a logistic regression model, a random forest model, and support vector machines with linear, polynomial, sigmoid, and radial basis function kernels. We are running them on a normalized subsection of our training data. The following is a table of the current AUC/ROC assessments for each of our models. Our goal for the rest of the project is to work on adding more features and refining our regularization terms for each model to improve our performance.

AUC Values using 10-Fold Validation:

Model	**AUC**
Logistic Regression	0.74
Random Forest	0.67
Linear SVM	0.678
RBF SVM	0.5
Sigmoid SVM	0.5
Polynomial SVM	0.660
MLP Classifier	0.67

Figures 1 and 2 show the ROC curve and AUC values for the logistic regression and random forest models.

Logistic Regression:

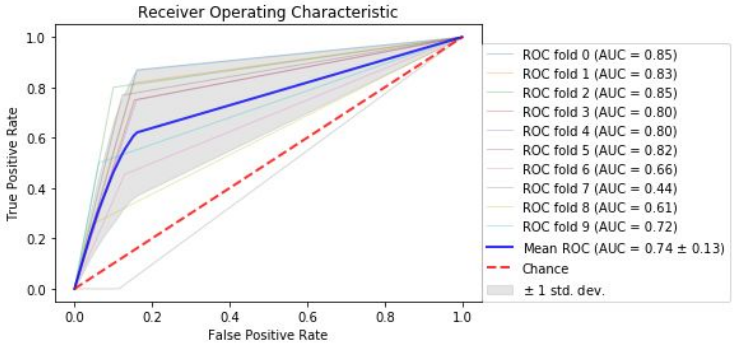

Figure 1

Random Forest:

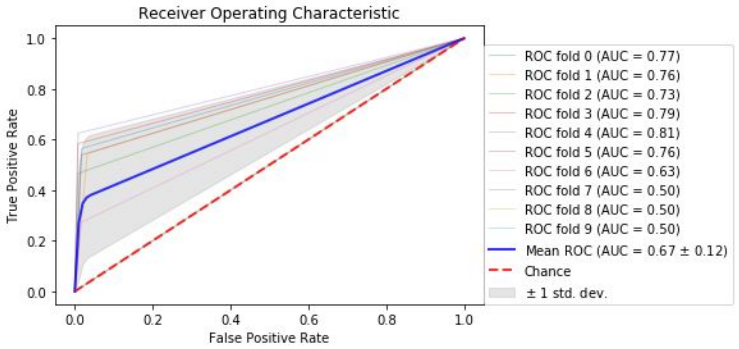

Figure 2

6. Timeline and Roles

Task	**Deadline**	**Lead**
Data Preparation	11/18/18	Nick
Implement Logistic Regression	11/26/18	Jacob
Implement SVM	11/26/18	Heather
Implement Neural Network	11/26/18	Jessica
Validation/Analysis	throughout	Nick
Prepare Project Update	11/29/18	all
Prepare report and presentation	12/11/18	all

References

1) *A. Alaeddini, K. Yang, P. Reeves, and C. Reddy. A hybrid prediction model for no-shows and cancellations of outpatient appointments. IIE Transactions on Healthcare Systems Engineering, 5(1):14-32, 2015.*
2) *H. Kurasawa, K. Hayashi, A. Fujino, K. Takasugi, T. Haga, K. Waki, T. Noguchi, and K. Ohe. Machine-Learning-Based Prediction of a Missed Scheduled Clinical Appointment by Patients With Diabetes. Journal of Diabetes Science and Technology, 10(3):730–736, 2016.*