Predicting Appointment Cancellations Using Patient Engagement Patterns

Heather Johnson, Jacob Kozol, Nicholas Pearce, Jessica Sandler
Group 6

Introduction

Even with reminders, some patients are unable to attend their appointments. When patients cancel their appointments on short notice or do not show up for their procedure at all, it leaves vacancies in hospital schedules that could otherwise be used to serve other patients. The goal of our project is to help hospitals predict in advance which patients will not appear at their appointments, so they can use that space to accommodate other patients waiting for appointments.

Dataset

Our dataset has 1997 training examples and 222 test examples. Each example includes an index, hospital ID, patient ID, registration date, procedure date, and whether or not the patient appeared at their appointment. We also have Medumo app engagement data for each patient, as well as some basic demographic and enrollment information for each patient (gender, date of birth, registration for email and SMS notifications). We use this data to create features for the month of the appointment date, the day of the week that the appointment falls on, the number of days between registration and appointment dates, the patient's age, the number of completed modules of each module type, and the number of messages received of each message type.

Cancellations within three days of colonoscopy appointment and no-shows are combined and labelled as 1. The procedures that took place are labelled as 0.

Patients' age and gender both have missing values at a rate of 46.9% and 29.8% respectively. We fill the missing values with the mean age and gender rather than dropping the entries altogether.

The training data was very imbalanced, with ~90% of the data points flagged as 0 and only ~10% of the data points flagged as 1. This meant that before we handled this problem, we were able to achieve high accuracies while obtaining low precision and recall. In order to handle this issue, we had to decide between oversampling the minority class (1) and undersampling the majority class (0). Because oversampling would have led to overfitting to the dataset, we have decided to accept the information loss associated with undersampling.

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. These studies led us to conclude that a logistic regression classifier would be a logical binary classification model to start with, and that cross validation would be an effective analysis tool.

Approach

We use binary logistic regression to separate the patients into the categories of took place (0) or missed appointment (1) with L2-norm regularization to counteract potential overfitting. We also use the newton conjugate gradient algorithm to optimize our solution. Aside from logistic regression, we also created three support vector machines with linear, sigmoid, and radial basis function kernels, a random forest classifier, and a binary neural network classifier.

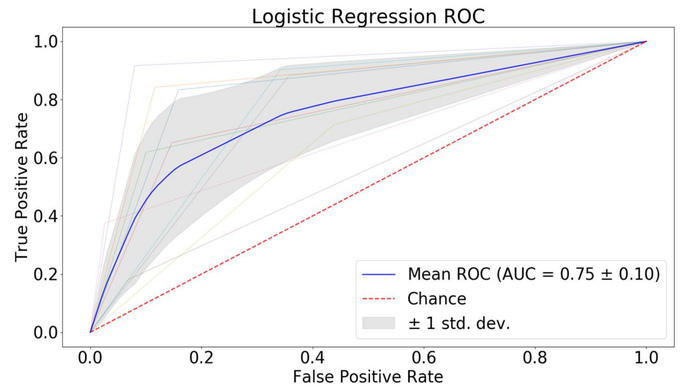
Logistic regression works well if there is a clear decision boundary, whereas a random forest performs well on non-linear features and decision trees perform well even with irrelevant features. A logistic regression classifier has a lower likelihood of overfitting in comparison with a random forest, especially since we use L2-norm regularization. The neural network can handle a larger feature space than a logistic regression, so as we added features it began to outperform the logistic regression in analysis using only training data. However, because neural networks are more likely to overfit to a dataset, logistic regression performs better on the test set.

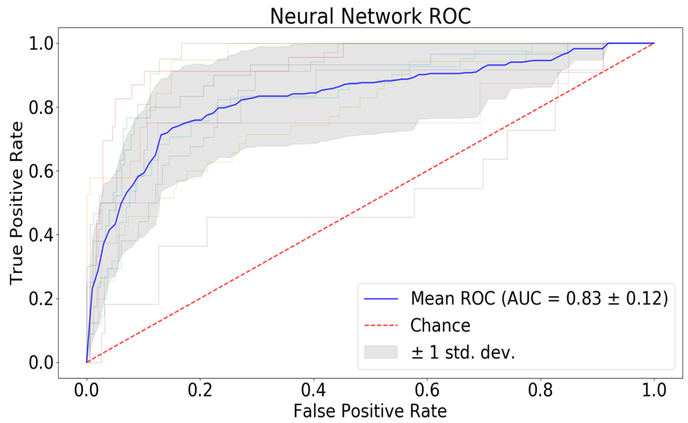
Analysis

The primary metric we use for analysis of the models is the area under the curve (AUC) of the receiver operating characteristic (ROC), which plots the true positive rate (TPR) against the false positive rate (FPR) of the binary classifications at varying threshold levels. We use this metric in conjunction with 10-fold cross-validation, which randomly splits the data into 10 folds and iteratively trains and tests models, holding out one fold as test data for each iteration. For each model, we compute the ROC and AUC for each of the 10 folds, then we compute the mean ROC and AUC across the 10 folds. Two graphs displaying the results of this analysis for the logistic regression and the neural network can be seen in the top right section of this poster.

The exclusive use of ROC AUC as a metric for performance initially led to an inflated sense of the robustness of the models. When the precision and recall of each model were analyzed, it was discovered that the high AUCs were the combined result of our models exclusively predicting the majority class (0) and the exceedingly high proportion of the majority class (~90%). To solve this issue, we balance the data by undersampling the majority class, as mentioned in our discussion of the data. This raised both precision and recall.

Classifier	Precision	Recall	Accuracy	ROC AUC
Logistic Regression	0.92	0.85	0.85	0.92
Random Forest	1.00	1.00	0.99	0.83
Linear SVM	0.95	0.96	0.96	0.80
RBF SVM	0.92	0.92	0.92	0.50
Sigmoid SVM	0.78	0.88	0.88	0.50
Neural Network	0.95	0.96	0.96	0.96





Discussion and Conclusion

The best classifier is the logistic regression model. The neural network and random forest perform the best on the training data with high precision, recall, and accuracy. However, the logistic regression model produces a higher ROC AUC than the random forest, and it performs better than the neural network on the test data. The neural network overfits for the training data, so even though it performs very well on that data, it drops off on the test data. On the other hand, the logistic regression model avoids overfitting at the cost of accuracy and recall.

If we continued to work on this project, we would split up the data into more classes defined by the time of cancellation (3 days before, 7 days before, etc.). We would expand our neural network to classify for multiple classes instead of just two, and create a clustering model, which may be better suited to a problem with multiple classes.

Our predictions may also be improved with more data, as we only had about 2000 data points for training. This would prevent overfitting to our dataset and overall improve the performance of our models.

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.