Predicting Delays in O.R.

BUDT-733
Fall 2007

Team# 7

Chris Low
Igor Nakshin
Xinyang Zou

December 11, 2007

To: Galit Shmueli, PhD.

From: Team 7

Re: Predicting Delays in O.R.

EXECUTIVE SUMMARY

According to Press Ganey's Physician's Office and Outpatient Pulse 2007[1], a widely used report in health care industry, on time appointment performance in ambulatory surgery is one of the most important factors with respect to patient satisfaction. However, how to best optimize the O.R. scheduling has been a long standing, highly debated issue within the healthcare industry.

Assisted by HMC, a regional medical center in northwest Kansas, we tried to utilize data mining methods to improve the performance of their O.R. scheduling. HMC has continually experienced frequent appointment time over-runs as the actual surgery often takes longer than booked by the hospital's scheduling software. As part of quality improvement effort, HMC desires to implement a report which would identify the surgery appointments that are likely to run longer than 15 minutes over the scheduled time. Using such a report, the O.R. personnel can optimize the daily schedule by moving appointments that are likely to have overruns to operating rooms with spare capacities during the day, thus reducing the bull-whip effect to the entire O.R operation.

HMC provided us with access to their data warehouse and a designated champion for the project in the hospital's decision support department. We also had access to Director of O.R. (through decision support department) for necessary domain knowledge.

During the course of the project, we have extracted the data from the warehouse and analyzed over 3,400 scheduled surgeries that occurred from January to August of 2007 using a number of data mining techniques. In creating our model we examined those variables that are available before the appointment actually occurred and tried to design a model could be easily implemented by the hospital's current IT infrastructure. Preferential treatment was also given to methods that have low computational costs since the model will actually be put to use by the hospital.

To tackle a tremendous amount of noise and variance in the available data, we have developed a logistic regression model using three-predictors that are easily obtainable and imputable from the existing data. When implemented in a report as a simple formula, the model shows desirable results in identifying appointments that are likely to have over-run time more than 15 minutes. The overall accuracy of the model is 86%, and the model is expected to be able to classify appointments that are *not* likely to be longer than scheduled by 15 minutes with accuracy of 97%. Finally, the model is deemed easy to implement with low costs.

A follow-up plan has been developed. We plan to report the results to HMC and, during the next several weeks, aid hospital's decision support champion of our project in implementation of the model.

[1] http://www.pressganey.com/galleries/default-file/outpatient-report.pdf

TECHNICAL SUMMARY

The goal of our task was predictive. In order to accomplish the goal, we created an extract from the hospital's data warehouse that contained surgical data from 2005 to 2007. Due to XLMiner's technical limitation we later trimmed the data set to ~3,500 cases between January and August 2007. Exhibit 1 displays the variables that were extracted. Because of the predictive goal the only variables that could be used in the model were those that are available *before* the appointment occurred.

Preparation

Prior to proceeding to analysis, we communicated with hospital's O.R. personnel through our decision support champion to inquire about what they think the possible explanations of why a surgery appointment would take longer than expected. The overwhelming majority response was that "nothing really matters other than the past performance of the surgeon and the surgery". Part of our goal was to confirm or deny this, therefore several patient-related variables (gender, race, age), surgeon-related variables (group, age, gender) and time-related variables (day of week, hour) were heavily used in data explorations.

Data Source & Exploration

The main data set selection criteria was HadAppt=TRUE. This criterion excludes the patients admitted to surgery from emergency room, which is an acceptable exclusion due to unpredictable nature of emergency patient's condition and surgeon's availability at the time.

The quality of the data is deemed highly accurate due to the processes and procedures in place in the hospital. However, there is a tremendous amount of noise and variance in the available data.

We spent a significant amount of time exploring the initial set or variables in various dimensions. Because of a large volume of categories in Serv, OpID and Surgeon variables, an early attempt to use dummy variables for these predictors yield very little utility.

Grouping of these three variables using pivot tables with respect to ActDur, PlanDur or LongerMin also failed to produce meaningful results. Same fate met other variables and transformations of them. Data plots, trees and PCA revealed little information in terms of explanatory relationship or behavior.

The summary statistics show that about 19.6% of the appointments took over 15 minutes longer than originally planned. Thus, the Naïve rule base error is 19.6% in our initial data set.

We believe this initial result is due to the following factors

- Explanatory power of single predictor is significantly weaken due to the complexity of each surgery
- Our initial data set is indeed very random with little modeling utility.

A Second Approach

The failure in initial exploration of data gave us enough reasons to believe that the O.R. staff may be right in their opinion about predictors. We went back to the data warehouse and regenerated the data set by including the predictors of SurgLatePct (percentage of surgeries when surgeon was over the expected appointment length by 15 minutes during last 6 month) and OpLatePct (percentage of operations that were over the expected appointment length by 15 minutes during last 6 months). These variables enabled us to build a logit-2 model with overall accuracy of 81.58% on test data and lift of 6.02% over Naïve classifier (see exhibit 2). Using the same predictors Discriminant Analysis, our second candidate, yielded slightly worse results than logistic model, and, therefore was rejected.

While the logit-2 model provided us with acceptable results with respect to our goal and with positive lift over the naïve classifier, it did not fully meet the secondary goal of being able easily implement the model in Microsoft Reporting Services report. The calculation of historical percentages of surgeon and operation lateness over 15 minutes proved to be very expensive with respect to functionality available in this software tool.

Additionally, the variables did not fully capture a *particular* surgeon's performance on a *particular* operation; therefore the objective model validity was not going to be easy to explain to the hospital's administrators.

Final Approach

Taking this into consideration, we returned to the data warehouse once again and regenerated the data set by including 90-day moving average of minutes, standard deviation and count with respect to particular surgeon and particular procedure combined. To clarify, for each appointment, we computed these three metrics by querying the appointments that occurred within 90 days of the "scheduled date" of the appointment in question. 90-day threshold was chosen after inquiring our decision support champion about the "relevant time horizon" of past performance.

After exploration of these new variables together with existing variables, the following parameters were found to be the best when constructing a predictive model:

1. PlanDur – planned duration of the appointment in minutes
2. MultipleProcs_Y – 1 if appointment has multiple operations, 0 otherwise
3. AvgLength – 90 day moving average of number of minutes same surgery took when it was done by same surgeon

The resulting logit-3 model had predictive accuracy of 83.7% on test data with 28.57% sensitivity, 96.24% specificity and 17.2% lift at 50% cut-off. We opted for higher specificity over sensitivity because we want the model to be more conservative by not aggressively identifying those appointments that are less likely to be longer by 15 minutes and thus resulting unnecessary actions by the O.R. staff.

Conclusion & Evaluation

- Reasonableness

The identified predictors make good sense. It appears very reasonable that over time of a surgery is dependent on a surgeon's the past performance, the complexity of operation and the scheduled duration. As compared to the other models we attempted in our initial data exploring stage, this model is much more parsimonious.

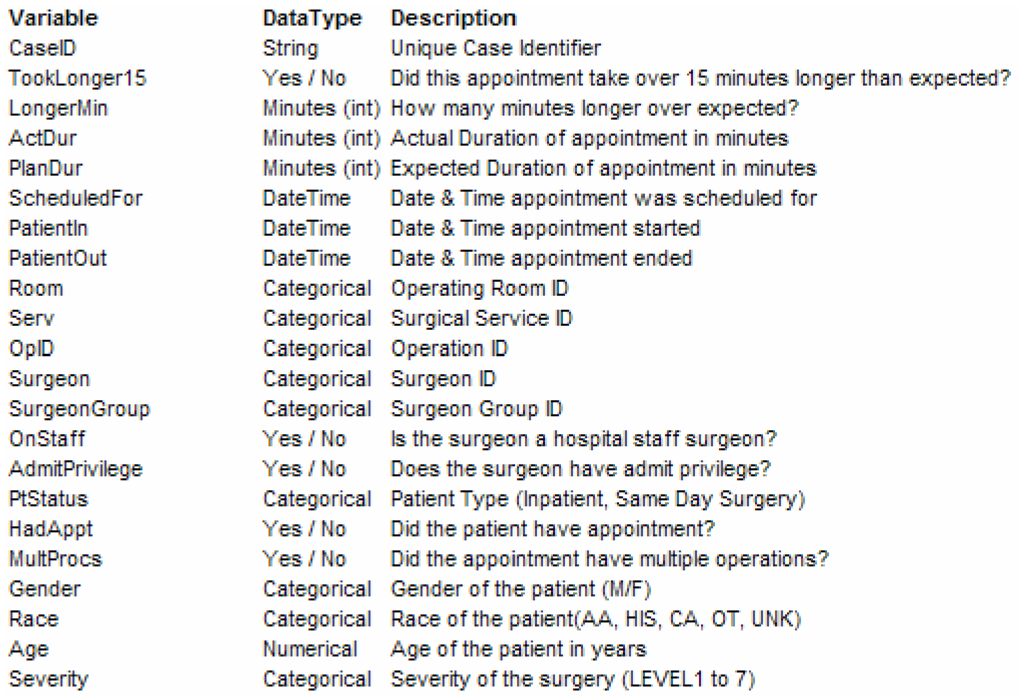
- Accuracy

The logit-3 model exhibits better performance than our second candidate logit-2 model. P-values of predictors are statistically significant. Although the model's explanatory utility is low (a low Multiple R-square), its predicative power is deemed satisfactory.

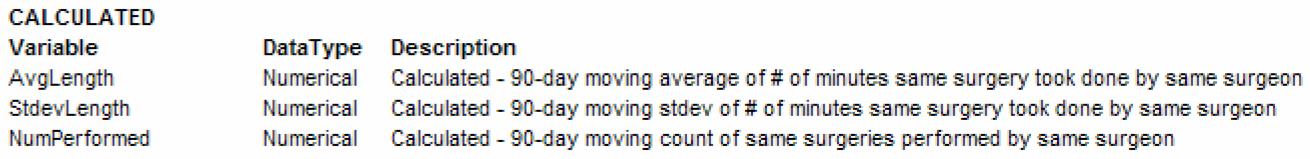
- Cost Consideration

The logit-3 model is much easier to implement in the Reporting Services environment as for each row of the data it requires a simple calculation of average appointment length during the prior 90 days filtered on the same operation and same surgery. It is the most cost efficient models we came across during our research.

Exhibit 1: Variables



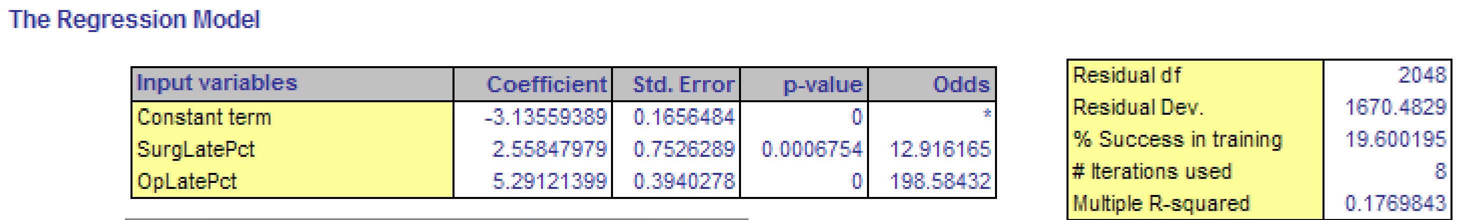
Variable	DataType	Description
CaseID	String	Unique Case Identifier
TookLonger15	Yes / No	Did this appointment take over 15 minutes longer than expected?
LongerMin	Minutes (int)	How many minutes longer over expected?
ActDur	Minutes (int)	Actual Duration of appointment in minutes
PlanDur	Minutes (int)	Expected Duration of appointment in minutes
ScheduledFor	DateTime	Date & Time appointment was scheduled for
PatientIn	DateTime	Date & Time appointment started
PatientOut	DateTime	Date & Time appointment ended
Room	Categorical	Operating Room ID
Serv	Categorical	Surgical Service ID
OpID	Categorical	Operation ID
Surgeon	Categorical	Surgeon ID
SurgeonGroup	Categorical	Surgeon Group ID
OnStaff	Yes / No	Is the surgeon a hospital staff surgeon?
AdmitPrivilege	Yes / No	Does the surgeon have admit privilege?
PtStatus	Categorical	Patient Type (Inpatient, Same Day Surgery)
HadAppt	Yes / No	Did the patient have appointment?
MultProcs	Yes / No	Did the appointment have multiple operations?
Gender	Categorical	Gender of the patient (M/F)
Race	Categorical	Race of the patient(AA, HIS, CA, OT, UNK)
Age	Numerical	Age of the patient in years
Severity	Categorical	Severity of the surgery (LEVEL1 to 7)



CALCULATED

Variable	DataType	Description
AvgLength	Numerical	Calculated - 90-day moving average of # of minutes same surgery took done by same surgeon
StdevLength	Numerical	Calculated - 90-day moving stdev of # of minutes same surgery took done by same surgeon
NumPerformed	Numerical	Calculated - 90-day moving count of same surgeries performed by same surgeon

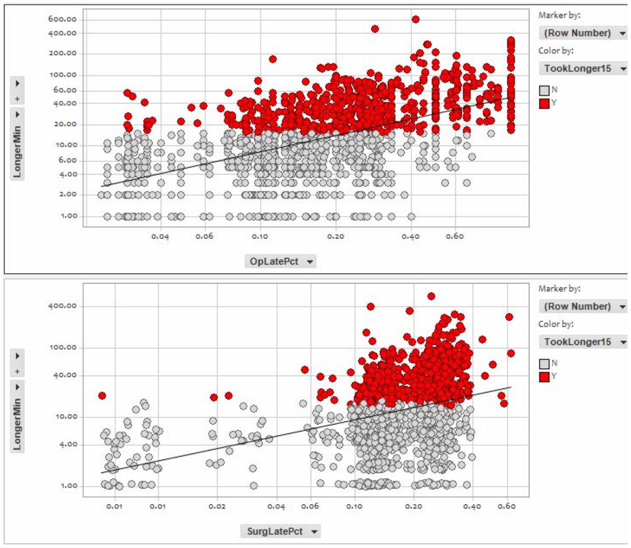
Exhibit 2: Initial logit-2

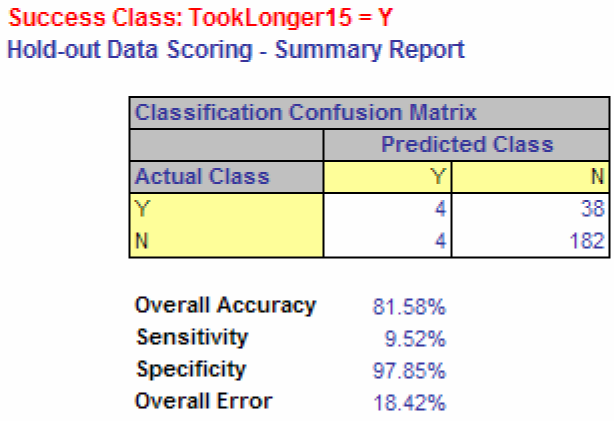


The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-3.13559389	0.1656484	0	*
SurgLatePct	2.55847979	0.7526289	0.0006754	12.916165
OpLatePct	5.29121399	0.3940278	0	198.58432

Residual df	2048
Residual Dev.	1670.4829
% Success in training	19.600195
# Iterations used	8
Multiple R-squared	0.1769843





Success Class: TookLonger15 = Y

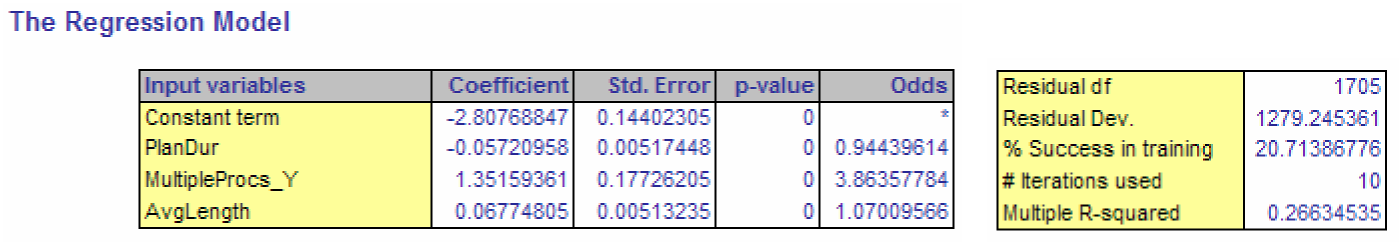
Hold-out Data Scoring - Summary Report

Classification Confusion Matrix

	Predicted Class	
Actual Class	Y	N
Y	4	38
N	4	182

Overall Accuracy	81.58%
Sensitivity	9.52%
Specificity	97.85%
Overall Error	18.42%

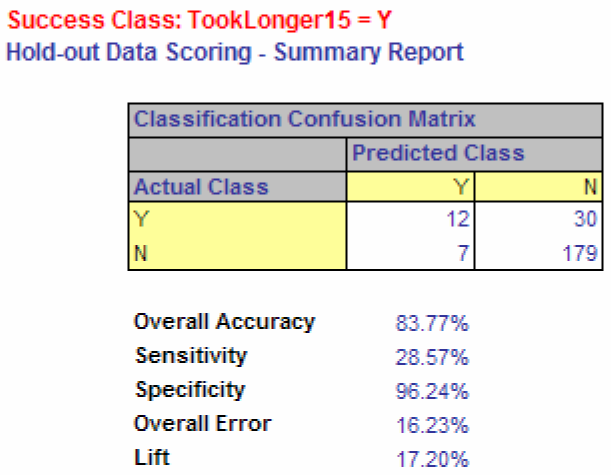
Exhibit 3: Final logit-3



The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.80768847	0.14402305	0	*
PlanDur	-0.05720958	0.00517448	0	0.94439614
MultipleProcs_Y	1.35159361	0.17726205	0	3.86357784
AvgLength	0.06774805	0.00513235	0	1.07009566

Residual df	1705
Residual Dev.	1279.245361
% Success in training	20.71386776
# Iterations used	10
Multiple R-squared	0.26634535

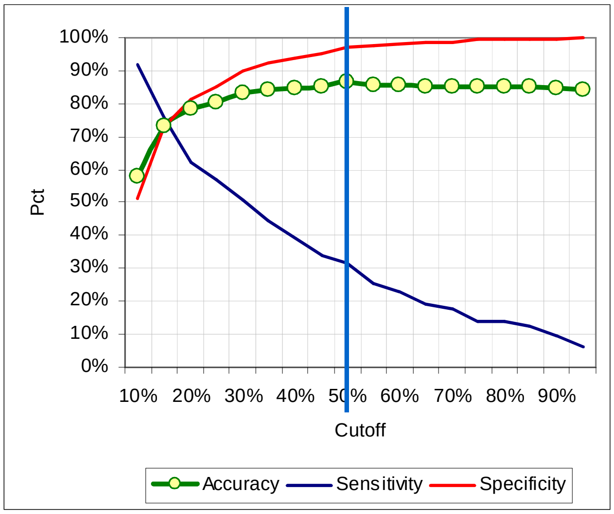


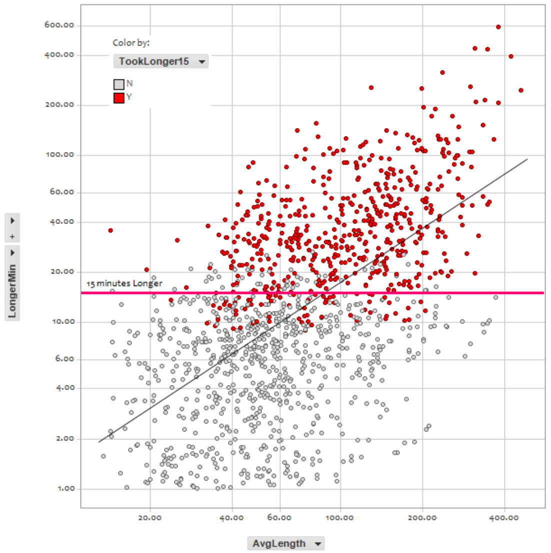
Success Class: TookLonger15 = Y

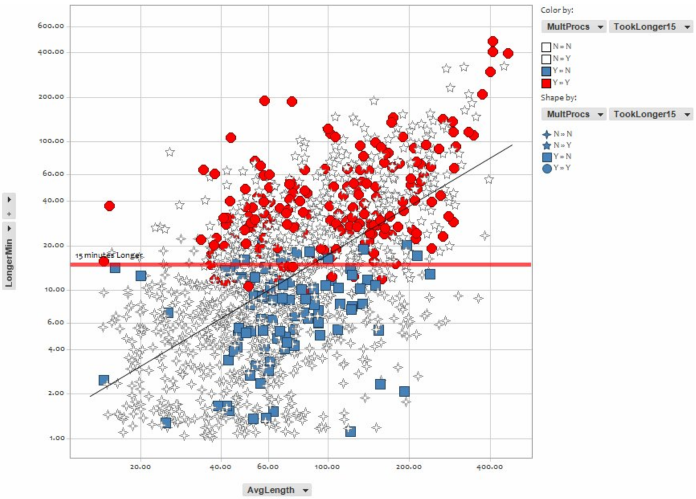
Hold-out Data Scoring - Summary Report

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Y	N
Y	12	30
N	7	179

Overall Accuracy	83.77%
Sensitivity	28.57%
Specificity	96.24%
Overall Error	16.23%
Lift	17.20%


Pct
Cutoff
Accuracy
Sensitivity
Specificity


Color by:
TookLonger15
N
Y
15 minutes Longer
LongerMin
AvgLength


15 minutes Longer
LongerMin
AvgLength
Color by:
MultProcs
TookLonger15
N = N
N = Y
Y = N
Y = Y
Shape by:
MultProcs
TookLonger15
N = N
N = Y
Y = N
Y = Y