Business Analytics using Data Mining

Project Report

Optimizing Operation Room Utilization by Predicting Surgery Duration

Project Team 4

102034606 WU, CHOU-CHUN

103078508 CHEN, LI-CHAN

102077503 LI, DAI-SIN

Advisor: Galit Shmueli

Executive Summary

Business Background and Motivation

Our main client is the newly elected director of the hospital who found that they are losing money in the operating room department due to the improper management on the surgery scheduling. Thus, he arranged a special committee to solve this hard task.

The business problem we face is that for surgery scheduling it is often subjectively assigned by the operating doctors. For the doctors, each of them has different preference of time schedule and even the operation rooms which results in the difficulty of proper scheduling. Our contribution is by accurately predicting the duration to help improve the existing scheduling system both from the space utilization and the human resource allocation perspective.

We are able to give practical advice based on the numerical evidence. Under the premise of patients' safety and the quality of the surgery, the medical team can arrange their time resource efficiently and the patients can be aware of the potential duration lasting intervals.

Source of Information

The raw data is the electronic version of the surgery agenda which is authorized by the administration department and is handled by the information department due to the privacy protection protocol. The data includes surgery records from a hospital during the years 2010 and 2011.

Analytics Solution

Once the reservation begins, the required predictors can be gained and will be plug in the specific model to predict the duration under different combination of variables. Based on the best interests of the patients, the scheduling can avoid extreme cases involving dramatic delay of the surgeries due to improper surgery arrangements. The analysis can also improve the utility of each individual operation room to prevent from idling.

The data mining project encounters the variable selection difficulty which is solved by using the Lasso Regression and Tree methods to select important variables. By comparing the existing method (use average surgery duration of individual division) and the proposed method, our method outperformed the naïve method in predictive power. For better interpretation, the rules generated by trees are also provided for the administration department to have a general concept of time duration given the diagnosis and surgery type.

Recommendations

The prediction result can provide the scheduling team to build up a reliable appointment in the system. By inserting the required variables the model will reveal the predicted output. With this outcome, the people in charge can better arrange in the situation dealing with conflict of multiple surgeries and availability of rescheduling. In the future, the system can integrate with the shift arrangement system and the accounting system to better control the operation expense and reduce the possible waste of medical resource. All affected members can respond instantly and give improvement advice on a regular basis.

If we divided the operation time into intervals, the only time that can be regulate will be the waiting time before a patient enters the surgery room. By controlling the waiting time to within 15 minutes, the operation time can be reduced by 5%. This result implies that the potential overtime expense can be reduced and the allocation of human resource can be more flexible. As for the scheduling rule, according to the simulation result using Arena (simulation software), by properly allocating the surgery room, the capacity can be increased by 7%. It shows that (instead of sticking to one specific operation room) if the surgery is scheduled in any available room, the cycle time can be significantly reduced.

*From the simulation result, the expected number of trials that can be performed is 5352 comparing with the original setting 5004 is a 7% gain in the capacity

Business Goal and Humanistic Evaluation

One major problem of the improper scheduling in the hospital is that there is uneven usage of the operation rooms. Our main purpose is to help the special committee to build a reliable scheduling procedure while maintaining the expense and the quality at the same time. Our business goal is to optimize the operating room usage including the establishment of an efficient surgery scheduling and enhancing the usage efficiency of both the operating room and the medical personnel.

To conduct a healthcare center in Taiwan, the maintenance of the operating room cost is one of the major issues. The expenditure of the operating room accounts for a significant proportion of the aggregate fee. Managing the operating room is not only a considerable spending, but also affects the quality of medical care and the standards to the hospital.

The dominating medical waste is result from the operating time delay and the uneven distribution to the operating room usage. Moreover, these circumstances can cause surgery scheduling difficultly and redundant payment on the overtime cost. Our project can be served as an effective way for reducing medical resources waste.

Analytics/Data Mining Goal

Our data mining goal is to predict the total operation time of a surgery begin with a patient enters the waiting room until he/she is moved into the recovery room. After we predict the entire operation time precisely for a new patient, we will be capable of setting up the arrangements to optimize the using of the operation rooms. Furthermore, we are going to build a supervised model to achieve the goal that we desire. For example, after plugging the information listed on the reservation notice, we will obtain the estimated operation time which we can better utilize in the scheduling.

The implementations of the data mining results do not involve the establishment of the standard time. The usage of the outcome lies in the arrangement of multiple surgeries in the same operations room or from different departments. With the prediction result the medical staff can make better decision on the surgery handling. Also, the prediction can be adopted to solve the uneven distribution of room usage.

Data description

There are 19 columns (variables) and 6783 rows (records without the title) within the raw data set. The nineteen variables can be sorted into five major categories. The first one is related to time measurement including Date, Holding Time, Room Time, Anesthesia Time, Start Time, End Time, and Recovery Time. The second is related to the staff involving in the surgery including Doctor, Assistant, Scrubbing Nurse, and Circulating Nurse. The third is about the surgery category including Division, Surgery Type, Anesthesia Type, and Diagnosis. The fourth one is about the serial number including Room#, Bed#, and Medical#. The fifth one is the description of the patient including Age, and Gender.

For those categorical data, we will use dummy variables to represent them. For some columns especially the second type dealing with the staff list, we need to apply the split function to the column and create new variables since the text may consist of multiple records. As for the dependent variable, we will primary use the value, Recovery Time minus Holding Time, as our response variable y.

Holding Time is the time a patient arrives at the waiting room and Recovery Time is the time when a patient arrives at the recovery room (a bed or the rest room). Thus, the operation time is defined as the duration between these two time measurements.

Table 1: Summary Table of the Variables

Numerical	
(Y)Operation Time	in minutes
Age	in years
Categorical (levels)	
Room#	20
Division	11

Gender	2
TreatType	315
SurgeryType	2
AnesthesiaType	12
Doctor	43
Assistant1	26
Assistant2	10
Diagnosis1	744
Diagnosis2	406
Method1	470
Method2	275
ScrubbingNurse1	26
ScrubbingNurse2	26
CirculatingNurse1	20
CirculatingNurse2	15

5 samples of variables

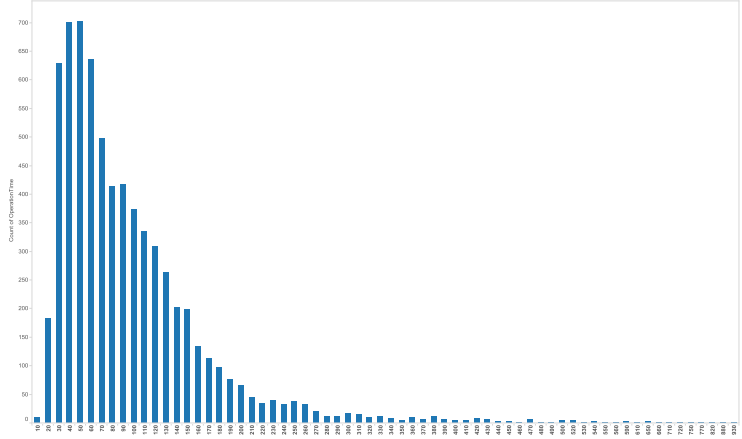
Date	Room#	Division	Age	Gender
099/10/01	9	Plastic Surgery	31	F
099/10/01	1	Orthopedics	62	F
099/10/01	3	General Surgery	49	F
099/10/01	9	Plastic Surgery	59	F
099/10/01	9	General Surgery	53	F

5 samples of variables

Bed#	HoldingTime	RecoveryTime	SurgeryType
Registration	08:15	09:05	Routine
NA	08:55	11:10	Routine
NA	09:05	10:00	Routine
Registration	09:20	09:55	Routine
Registration	09:30	10:20	Routine

*Bed#: Registration means a patient does not need a bed and need to return to the doctor for an appointment; NA means the patient does not need a bed in the recovery room.

Graph1: Histogram of total Operating Time (X-axis: minutes Y-axis: count)



Data mining solution

At first, we explored and pre-processed the data using available domain knowledge. For data exploration, we used software to visualize the datasets. Then, we generated dummy from each level of the variable and use dummy variables to present different combination of the datasets, for example the division in the hospital, the anesthesia type, surgery room and surgery type, and emergency or routine. We also tried to group together those categories that have similar surgery time boxplots and binned these variables to do the dimensional reduction. As suggested by the advisor, we binned the variables with many levels according to the dependent variable and found that the variables can be reduced to 550. At the meantime, we found out some characteristics from the binning result. As for the next step, what we should do is the variable selection through Lasso Regression, Regression Tree and Random Forest. By doing the variable selection, we exclude the useless measurements and start to train our model by several supervised methods. We primarily use general linear regression for explanatory purpose. At last, we use the result of Naive Bayes which round up the surgery time into a 5 minutes interval to compare with other models.

Table 2: RMSE of Training Data

Lasso	NB	RTree	RForest
30.723	63.502	44.322	2.047

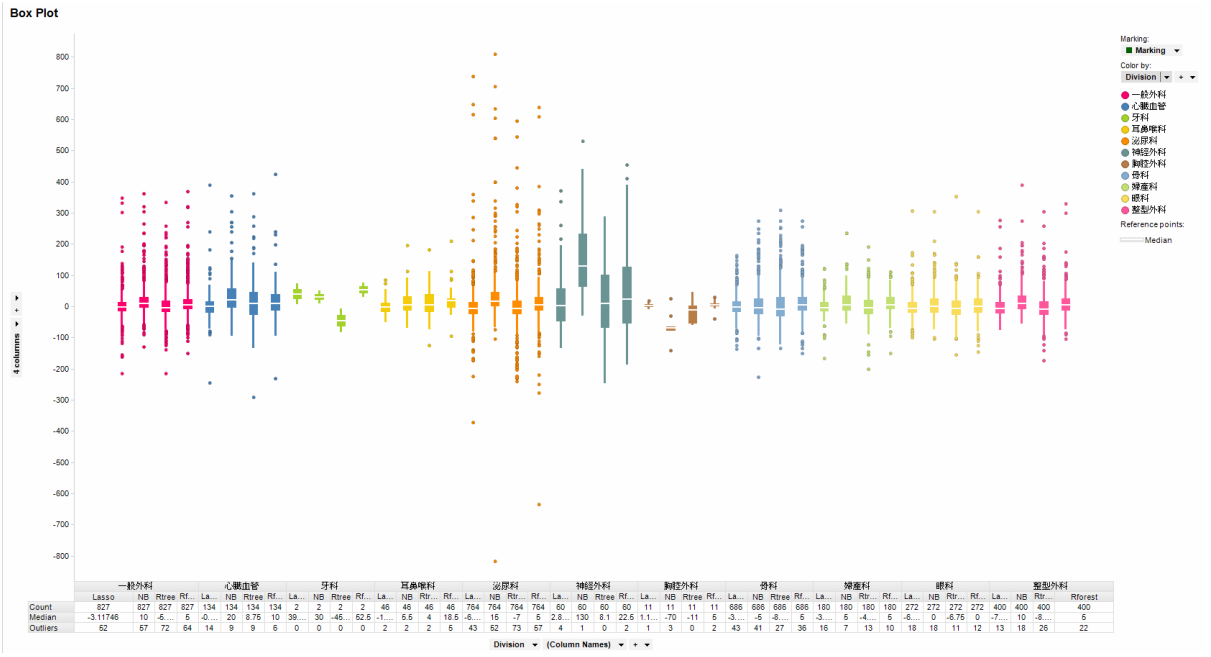
Table 3: RMSE of Testing Data

Lasso	NB	RTree	RForest	Naive
47.353	54.509	54.898	55.364	77.130

Table 4: MAPE of Testing Data

Lasso	NB	RTree	RForest
0.322	0.396	0.448	0.335

Graph2: Boxplot of the residuals from each methods divided by individual division



From the result of random forest, we can see that the model over fits the data from the training set. From table 3, we can tell that our proposed methods are all better than the naïve method which is using the average time for prediction. The criteria for MAPE is that for an acceptable model to be less than 0.5 and for a good model to be less than 0.3 which suggests us that the Lasso Regression should be an ideal choice for prediction purpose. From graph 2, we can see the model performance by each division and we can conclude that the prediction of Urology (the orange one) operation time is more challenging. Since we did not include the severity of the diseases, even the predictors are similar, the diversity of individual patient is really high. The medical staff gave an explanation that for some surgeries, partial ultrasound doesn't provide enough information as whole body X-ray scanning.

Conclusions

It is costly to underestimate the operation time which may lead to possible overtime payment. Thus, from this perspective, the linear model should be an ideal choice. As for the practical usage on scheduling, we will generate some rules from the trees for further interpretation.

The model can be run in real-time and we have used the existing data to build up a predictive model. The new data can be collected for testing and prediction. The error rate will be calculated to set up a rolling mechanism for reconstructing the model.

The model we built can be adopted to optimize the operating room usage rate through predicting the total operation time and also used in setting up the scheduling system. As for the predicted value, the confidence interval will be provided for better accuracy. Other hospitals can refer to this result to take advantage of our proposed method to construct their system. For future research, we can collect more measurements, such as the cause of the surgery, the level of the disease and the working schedule of the operating doctor for future model construction.

Appendix

Since the variables contain private information from individual patient and the delay time of individual surgery, the full table will not be provided.

Table 5: 20 variable selection by Lasso Regression

"variables"	"c.inds."
"1"	"(Intercept)" 90.9633522648004
"2"	"V2" 0.0584711980744214
"3"	"V3" -13.487857679418
"4"	"V7" 2.70601364381187
"5"	"V11" -7.89434829778163
"6"	"V23" -3.96245937153829
"7"	"V24" 5.06378480047977
"8"	"V28" 34.2288505518762
"9"	"V32" 11.3448136168794
"10"	"V39" -6.98947496590069
"11"	"V43" -4.69839089977925
"12"	"V48" -28.1433919627335
"13"	"V58" -2.73632169828727
"14"	"V61" 0.81244091094322
"15"	"V73" -0.859546428139521
"16"	"V86" -11.9268365908761
"17"	"V88" -1.77954266050739
"18"	"V92" 47.1470815566583
"19"	"V94" 16.4025388227867
"20"	"V98" -7.39266806393305

Graph3: Variable selection by Regression Tree

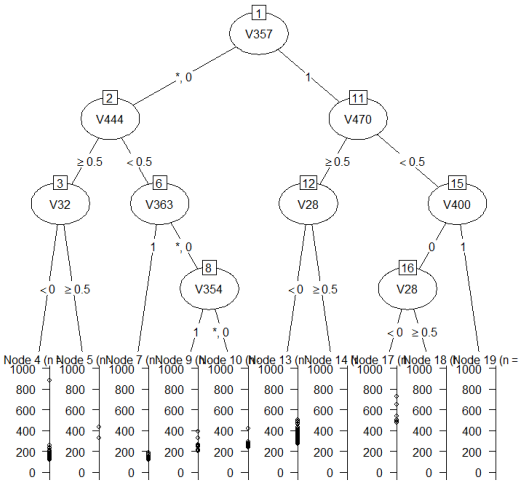


Table 6: Comparison Result of Full Linear Model and Lasso Regression

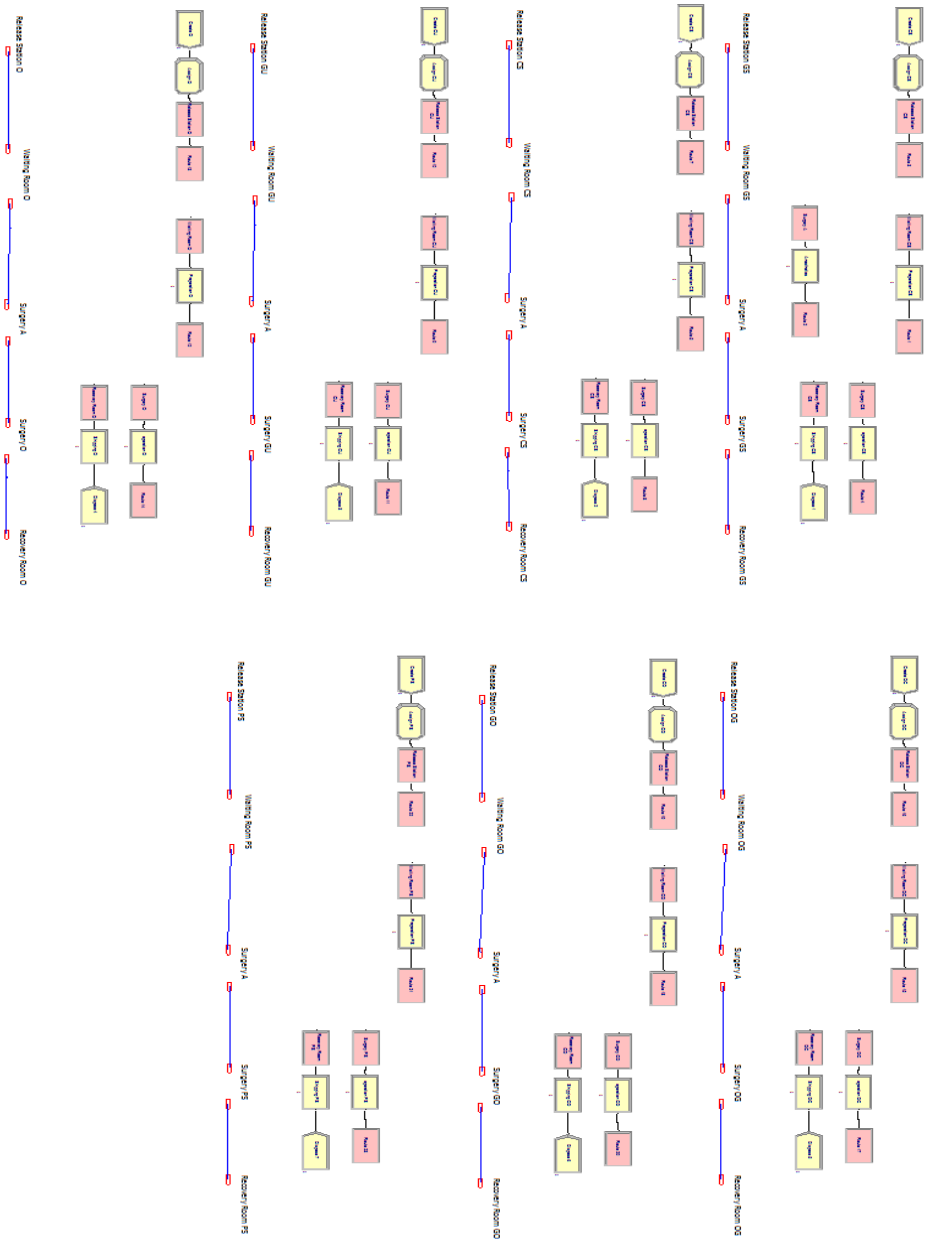
	Full	Lasso
variables	2424	699
R-Sq	86.73%	83.87%
R-Sq(adj)	80.41%	82.14%

The full model indicates the result of linear model without variable selection (with all inputs).

	Naïve	Full	Lasso
RMSE	73.75502	66.25416	43.36949

The naïve method is our benchmark which is equal to using the average value for prediction. The full model doesn't really outperform the benchmark that indicates the possibility of over-fitting.

Graph4: Setting of Arena (simulation software)



The surgery procedure can be divided into: Waiting Room → Anesthesia (Surgery A) → Operation (Surgery B) → Recovery Room
We use the distribution fitting function with the real data to obtain the time distribution of each stage of a surgery:

Department	Arriving time	Waiting Time	Anesthetics	Surgery	Recovery
General	EXPO(2.5)	2 + WEIB(23.5, 1.38)	0.999 + EXPO(9.72)	2 + LOGN(41.6, 67.1)	Constant(5)

(Unit in hours)

(1).Replication Length = 8 (hour)
(2).Number of Replications = 365 (days)