DUKE
THE FUQUA SCHOOL OF BUSINESS

Personalized Health Care Analytics

Background

The Health Care industry annual revenue exceeds a trillion dollars with over seven hundred thousand health care companies in the US. The size and scope of this massive industry generate a variety of opportunities and challenges. Part of these challenges arises from the unavoidable role Government needs to play when regulating public health policies. For example, how much to tax smoking to offset associated health care costs or who to target ad campaigns to reduce smoking more effectively. Such decisions are difficult and controversial and as such they require convincing evidence to be implemented. Other challenges arise from logistic issues with industry specific constraints. For example, hospital staffing has proven difficult because of the very nature of the service. One needs to balance continuous shifts and mandatory rest periods required by syndicate laws and safety practices. Fundamentally, the demand for health care has a high uncertainty due to emergencies and unforeseen events. Finally, even though more accurate tests and increasingly effective drugs are being developed, their costs have grown to unsustainable levels for many primary care facilities especially in rural areas. Therefore, in many areas, technology is no longer the bottleneck, rather finding cost effective solutions which requires compromises is the issue.

It is clear that health care analytics can play a significant role in these opportunities. Given the pressure to diminish health care costs and the size of the assets involved (from machines to personnel), even small gains in efficiency can translate into important contributions to the balance sheet.

Data Analytics to Achieve Cost Effective Screening: Birthweight Pilot Study

The use of health care analytics can provide more accurate and personalized assessments when patient specific information is available. However, analytics can be used as a cost effective screening option only if the required patient information is available at a low cost. To convince stakeholders typically a pilot study is important. Moreover, success in one application will typically lead to additional resources in the future.

We will consider screening for risk of complications with newborns at low cost (which is of interest in rural areas). In particular, we will exploit the fact that many complications are associated with low birthweight. Thus cost effective screening procedures are needed for that. The first step in the pilot study is to predict birth weights based on available information. Detailed “natality data” are recorded for nearly every live birth in the United States. Information on maternal characteristics (age, education, race, etc), prenatal care (number of prenatal visits, smoking status, etc) and birth outcomes (birthweight) is collected by each state (with federal guidelines on specific data-item requirements). Unfortunately, at this early stage of the pilot study, due to confidentiality restrictions, we do not have full access to the medical and demographic information of patients. Therefore, results obtained should be seen essentially as a “lower bound” on the predictive performance of this approach.

Exhibit 1 describes the data available that consists of 198,377 records of infants' and associated demographic information. The large sample size will allow us to achieve precise estimates and forecasts.

"Personalized" Staffing Decisions in Health Care Facilities

Historically, staffing decisions have been made to manage observed typical demand. This usually requires substantial adjustments and the placing of several medical and staff members on call. The use of data to improve staffing decisions is not new. For example, for many decades aggregate level of daily "patients' demand" has been recorded and staffing decisions typically adjust for the day of the week, hour of the day, and holidays. Nonetheless, every hospital department faces different types of demands. Indeed, cardiologists and dermatologists face substantially different uncertainties.

A complementary approach is to use individual specific information of the current patients. Indeed, in the vast majority of days, most duties are due to current patients whose detailed medical records are known and, to some extent, ready to be used. In principle, this can be used to forecast short and medium-term needs more accurately by conditioning on all the information currently available. This allows also for real-time updates on schedules and "on call" duties in a much more adaptive way.

Baseline and Deployment

In any data analytics application, we will be challenged on two counts that might not seem a part of the data analytics per se. Nonetheless, they are fundamental aspects of the creation of successful data analytic tools. The first one is the need to quantify the impact of the proposed methods achieve. That is, to establish a proper baseline for comparison. The second is the implementation of the proposed approach. Even the most promising data analytic tool can hit a wall if its deployment boils down to substantially changing the way an organization works.

References

Statistics Brain Research Institute, Health Care Industry Statistics.
http://www.statisticbrain.com/health-care-industry-statistics/

Exhibit 1 - Variable Descriptions

id – unique identifier for the customer
birmon – identifier for the shopping point of a given customer
weight – infant birthweight (in kilograms)
black – whether the mother is black or not (0=no, 1=yes)
married – whether the mother is married or not (0=no, 1=yes)
boy – whether the infant is a boy or not (0=no, 1=yes)
tri1 – first prenatal visit was in the first trimester of pregnancy (0=no, 1=yes)
tri2 – first prenatal visit was in the second trimester of pregnancy (0=no, 1=yes)
tri3 – first prenatal visit was in the third trimester of pregnancy (0=no, 1=yes)
tri.none – columns of zero (to be ignored)
novisit – did not perform a prenatal visit (0=no, 1=yes)
ed.hs - mother's education is exactly high school (0=no, 1=yes)
ed.smcol - mother has some college education but not completed (0=no, 1=yes)
ed.col – the mother has a college degree (0=no, 1=yes)
mom.age – mother's age
smoke – whether the mother smokes or not (0=no, 1=yes)
cigsper – number of cigarretes per day smoked by the mother
m.wtgain – mother's weight gain during whole pregnancy (pounds)
mom.age2 – mother's age squared (to allow a more flexible functional form)

Personalized Health Care Analytics Assignment

Instructions

This is a team assignment. Each member of the team receives the same grade. Submission is online (see course webpage). In order to be graded, you need to upload one pdf file (no longer than 3 pages with font size 12pt) and your R script (this should be well commented and run without errors). Any additional material you judge relevant that complements your submission can be submitted as additional files. Make sure that the section number and all names of the team members are clearly listed. Late submissions (but submitted before in-class discussions) or inappropriately formatted cases will have points deducted. Missed cases are worth 0 points. Important: submit a PDF file and follow the naming convention (see Syllabus).

Assignment

Your answers should be clear and provide unambiguous recommendations when asked. Please provide explanations for your answers and any outputs that you feel are needed to support your argument.

1. Before running any model, pick any two variables and attempt to show an (interesting) relation via visualization. This requires one to formulate a question, and to communicate clearly a conclusion based on data visualization. Ideally this would suggest ways to act on the issue. (For example, during pregnancy, does the amount of smoking seem to diminish with education level? This can provide guidance on where to place educational ads regarding smoking during pregnancies.)

2. Next consider the 10 dummy (binary) variables in Exhibit 1. Test independence among all 45 combinations. First using the traditional 0.05 rule. Second controlling the false discovery rate (conservative and FDR with q=0.001). Any discrepancies? If so, which one do you believe got it wrong this time?

3. In your opinion, can any of the variables provided in Exhibit 1 help to predict birthweight? Since opinions do not provide strong arguments, provide a simple evidence based on data.

4. Run a multiple regression with 14 of the variables described in Exhibit 1 (all except for **id, birmon, tri.none,** and **novisit**). Which variables are statistically significant? Apply the 0.05 cut-off rule and also control for false discovery rate (conservative and FDR with q=0.001).

The following questions do not require additional data analysis calculations.

5. In rural health care, primary care clinics cannot rely on expensive screening tests. To increase screening capabilities of detecting risky pregnancies early (end of first trimester), consider using the model in Question 4 to forecast birthweight since it is inexpensive. Discuss issues with implementation and choice variables.

6. Consider more broadly the various departments in a hospital which forecasts demand for services (with potentially additional information) that is used to build simulation software to more precisely set the number of nurses and other employees. Briefly discuss potential issues regarding deployment. Propose a suggestion that addresses at least one of them (even if only partially).