Optimizing healthcare analytics: How to choose the right predictive model

An EXL whitepaper

Written by

Lisa Chancellor
Vice President and Practice Head,
Healthcare Provider Analytics
lookdeeper@exlservice.com

Shivang Baijal
Business Analyst

Healthcare providers increasingly understand how predictive analytics can tackle serious industry challenges such as patient readmission and mortality rates. Less understood are the advanced techniques used to generate these insights. One example is the widely used **logistic regression** algorithm, frequently used for early intervention by targeting patients based on medical history (such as the probability that diabetes patients will end up in emergency room, or an asthmatic's readmission due to environmental triggers[1]).

There are myriad examples of predictive analytics in play today across the provider spectrum. ICUs and operating rooms are data rich environments. They use PDMS (patient database management system) that actually improve the quality of medical charting and can compute ICU risk prediction scores to provide positive impacts for a healthcare practitioner. Some healthcare providers even turn to analytics to analyze peak emergency room traffic flows to avoid bottlenecks in hospitals. Healthcare providers, also feeling the financial pinch of high 30-day readmission rates, often turn to analytics to keep patients home.

The use of predictive analytics in this space is prolific. This paper seeks to delve into understanding the nuances of each of these models and their practical applications. This will help guide the providers regarding the usage of different models in varied real-time applications to attain optimal results.

Conventional and advanced machine learning algorithms

Modeling techniques are broadly divided into two classes - segmentation and scoring. "Segmentation" is based on clustering patients with shared features and then selecting the cluster with maximum expected response according to the model. "Scoring" assigns individual patients a specific probability score to

turn up with a particular disease, and the targeting dataset can be tailored at each patient level[2].

Generally, logistic regression is used to develop the predictive analysis by computing the probability scores for the patients for a particular event. When the data gets non-linear, a decision tree algorithm is implemented that works on the principle of segmentation. Although prone to over-fitting, it provides an efficient solution if the further additional techniques of bagging and boosting are used. Rarely deployed machine learning algorithms like SVM, ANN (Artificial Neural network) can be used alternatively to provide an optimal solution.

The following section dwells on the technical nuances of each modeling framework.

Logistic regression

This classification technique was developed by David Cox in the year 1958. In general practice, two types of regression are used - linear and logistic regression. It is a special type of regression where the binary variable is predicted by finding the relation with the independent variable (categorical or continuous).The output of the logistic regression is the probability of a response occurring. But, there are certain assumptions that need to be taken regarding the probability distribution of the variables for logistic. The distribution of probability is assumed to be logarithmic.

The below equation is used to compute the probability of an event based on the input independent variables x1,x2,….xn.

$$\ln\left(\frac{f(x)}{1-f(x)}\right) = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots .. \beta nxn$$

Logistic regression can be used in health care profiling where factors like gender, age, BMI, systolic blood pressure, cholesterol, heart rate, diabetes, smoking, previous heart infarct are used to predict infarction; or in some cases prediction of undiagnosed diabetes using the postprandial time and random capillary plasma glucose[3]. Another application of logistic regression is leveraging the mammography database to predict breast cancer[4].

Decision tree

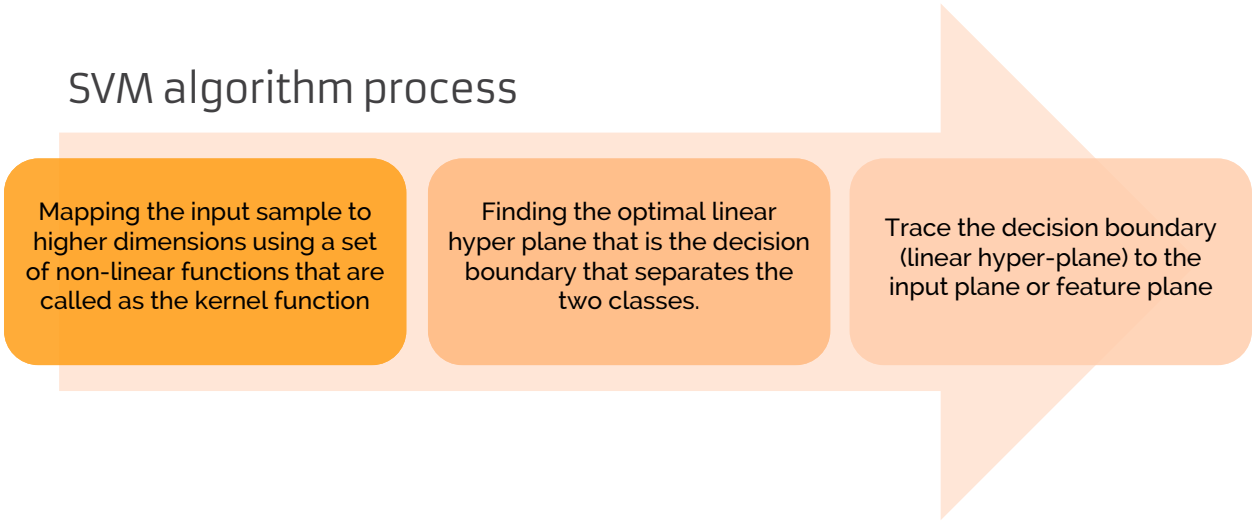
Decision tree classification works on breaking the dataset into smaller subsets on the basis of features (variables) and assigning the probability of an event. The tree consists of decision nodes where a subset breaks into further sub-sets of data and the leaf node is the final node that cannot be further broken down, which represents the classification. The various commonly used algorithms that are used to implement a decision tree are ID3, C4.5, CART, CHAID and MARS.

The main application of the decision tree is to aid the physicians' decision based on the evidence-based medicine approach, generally used when the physicians face challenges in medical decisions or when the outcomes of the treatments or medicines are uncertain[5,6]. It can also be used to diagnose heart disease patients.

SVM (Support Vector Machine)

SVM is one of the complex machine learning algorithms used to classify objects. The algorithm follows mainly three steps.

SVM algorithm process



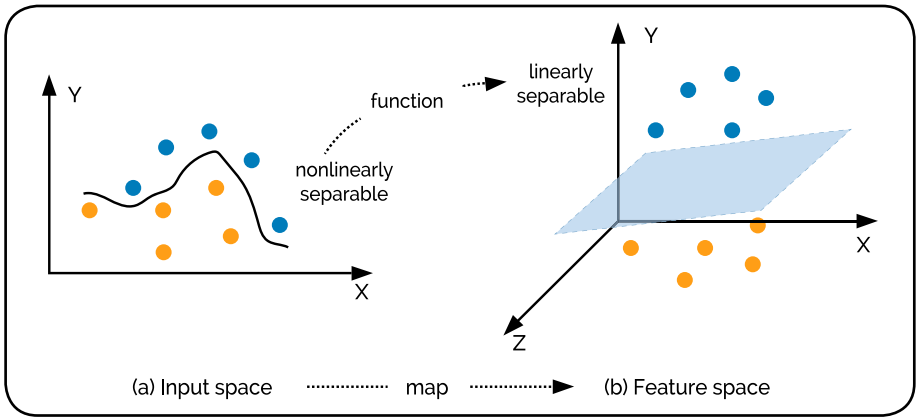


Figure 1

The boundary emulated might be non-linear as seen in Figure 1, in the feature plane, that makes SVM an excellent classifier in case of non-linear data.

It has been observed that SVM played an important role in designing an algorithm using a data from a self-reported questionnaire to predict the medical adherence for heart failure patients[7]. SVM is used for data mining and in prediction of common liver diseases, diabetes and pre-diabetes based on commonly used demographic and behavioral data[8].

Comparative case summary

The healthcare industry is witnessing a shift from volume-driven treatment to value-based care. To keep pace with this ongoing paradigm shift, healthcare providers are increasingly incentivized to leverage predictive analytics concurrently with healthcare techniques to optimize their performance in the market.

In this endeavor, providers are gradually roping in analytical intervention at different points of care to drive insights to improve population health, improve patient experience and reduce healthcare cost.

Providers are executing different campaigns like acquisition, retention, win-back, etc. to consolidate or expand the patient base.

However, without analytical support, providers can confront several issues while targeting patients for different service lines. Providers may incur huge sunk costs by targeting patients who are already present in their system (Commonly known as FAR –False Acceptance Rate); moreover, they can even miss out on potential revenue by missing the potential clients that could have been a success had they been targeted (Commonly known as FRR-False Rejection Rate).

Thus, targeting strategy for the above campaigns depends crucially on the results of predictive models. However, type of predictive analytics model applied changes when the nature of data changes.

Hence, an in-depth understanding of the data and its characteristics is important to select the optimal model in each scenario.

Considering the following two cases, data can be classified on the basis of the linearity of the independent variable (features that act as input in the model), if the data features are linear with respect to the dependent variable, then the logistic regression will suffice and further complex modeling algorithm in further prognosis of the target will result in wasting time and resources. However, non-linearity in data may demand using more complex models.

The below examples provide the different scenarios for a healthcare provider planning to launch a bariatric campaign facing different data types.

Example 1 – simple data (linear feature distribution)

In case of linear data, the use of logistic regression to predict whether a patient falls in the category of bariatric or non-bariatric will be the optimal method. In the given case, logistic regression will prove to be an excellent classifier and the accuracy of the model will be likely to bring good outcomes to the provider as highlighted in figure below.

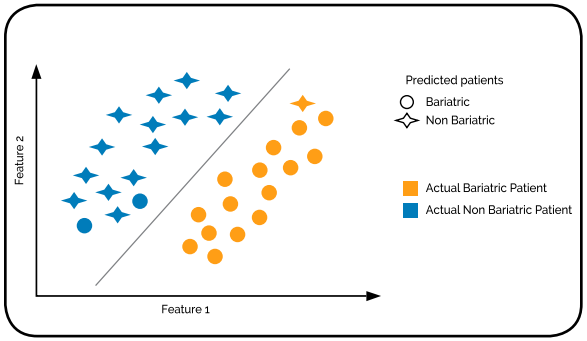


Figure 2

Example 2 – non linear data (complex approaches)

Figure 3 depicts the population divided into two categories of bariatric and non-bariatric population on the basis of two features. The aim of the classifier is to separate the population into the bariatric and non-bariatric in this more complex non-linear data.

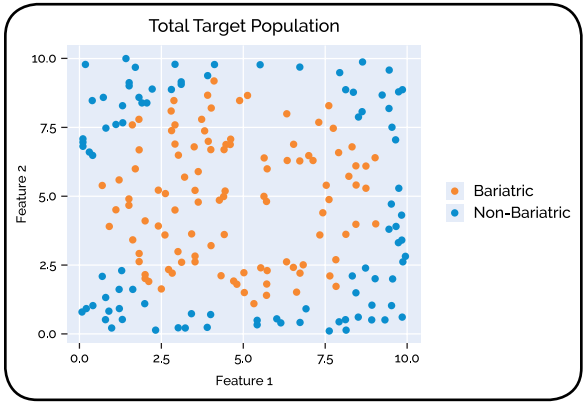


Figure 3

As we have only two features, the feature space will be two dimensional. The following diagrams depicts the application of logistic regression, decision tree and SVM on this data and tries to highlight the increasing efficiency in targeting moving from logistic regression to decision tree and finally to SVM.

Classification using the logistic regression

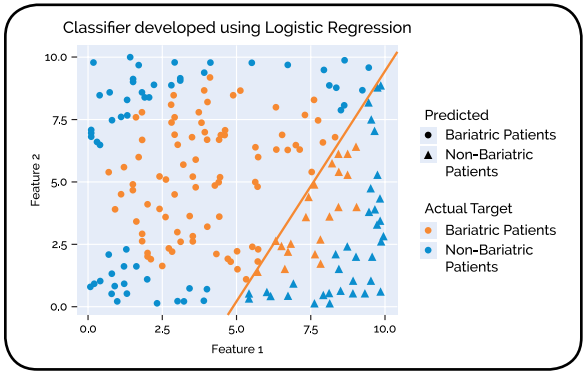


Figure 4

In logistic regression, generally a cut-off probability is decided upon that is assumed to be more than 50 or 60 percent, and the output that will be emulated by the training of the model will look like a straight line as shown in figure 4. The logistic regression has misclassified a lot of patients, as shown in the figure where it has kept circles as bariatric targets and the triangular as the non-bariatric target although the actual classification is depicted using orange and blue colors.

Classification using decision tree

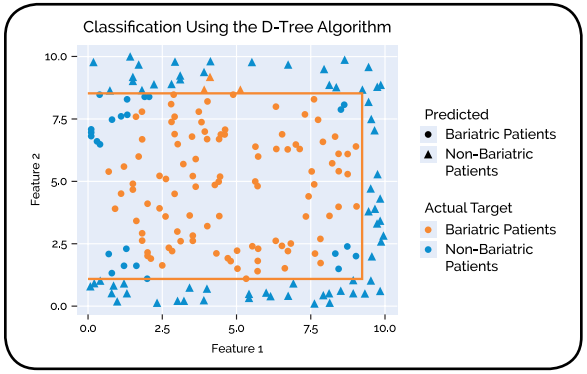


Figure 5

The decision tree will also emulate a decision boundary, the rule developed by the decision tree will divide the population in form of parallel lines (or cube or cuboids in case of high dimension features plane) as shown in figure 5. The targets within the orange boundary are predicted as bariatric target and outside are classified as non-bariatric target.

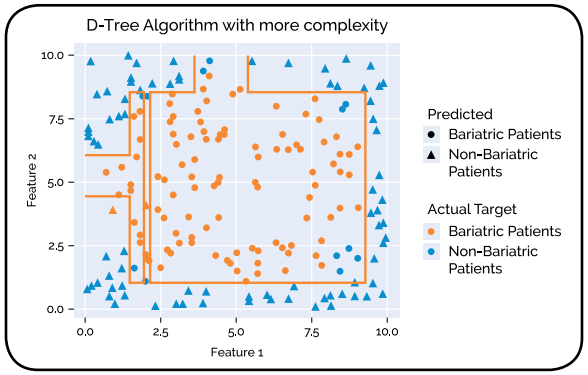


Figure 6

To increase the accuracy of classification and to reduce the misclassification rate requires making the D-tree complex. It will try to develop a rectangle that will divide the population as shown in figure 6. The only issue will be that model will be highly biased, heavy and unstable.

Classification using the support vector machine (SVM)

The SVM Algorithm's main step is to convert the feature space into a kernel space by adding extra dimensions to dataset until it finds a kernel space that can classify the target and non-target population. In this case, it actually added an extra dimension to the features and then developed a rectangular plane in kernel space that separated the population into bariatric and non-bariatric. The particular rectangular plane is a circular decision (non-linear) boundary as shown in figure 8 that accurately separates the two type of population.

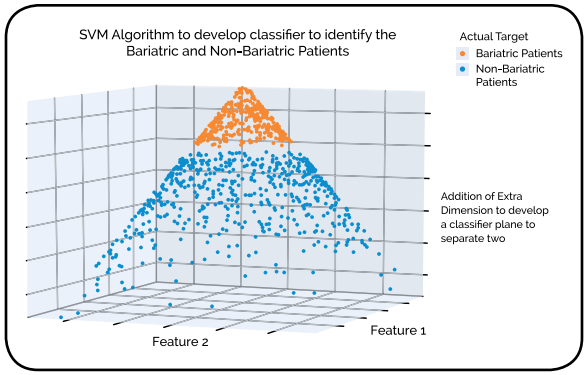


Figure 7

EXL look deeper.®

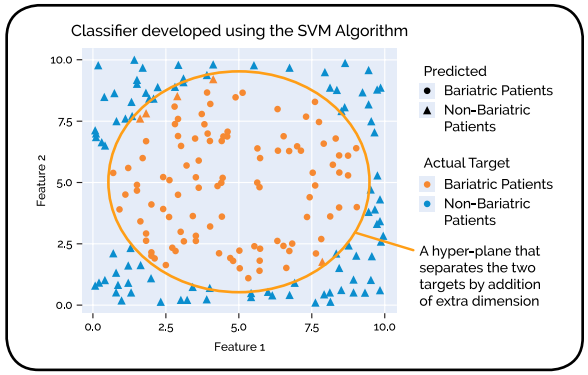


Figure 8

The SVM actually trained a rectangular plane boundary in the above shown kernel space, but that was a circular decision boundary in the features space. This helps SVM to develop a classifier that can also perform well on the non-linear data.

The SVM outperforms the other two because the nature of the data was nonlinear.

The result cannot be generalized as different data will have different methodology to develop a model. There are cases when the data model developed using logistic regression is more accurate than the SVM. It has been generally observed that the high-bias classifiers are accurate for small dataset and low-bias are best for big dataset.

Results

Although there are certain pros and cons of each and every technique, the main reason why logistic regression is mostly used in healthcare predictive analytics is because it is simple. The probability score and the cut-off of the probability score can be decided and hence the whole campaign can be tailored by the provider itself, but the factors like large number of features or non-linear features are serious challenges to logistic regression[9].

The decision tree algorithm has the ability to perform well even for the non-linear features and select the most important variables accounting for the variable interaction. But, the main challenge decision tree faces is that it develops models that are highly biased to the training dataset, and as the complexity increases the tree becomes heavy, unstable and prone to over fitting. Unlike logistic regression, it does not provide the individual probability score.

In case of non-linear data, SVM outperforms both the decision tree and logistic regression as it can handle high feature space. It also generates its own SVM score that can be converted into

probability score using the Platt scaling. The only challenge to develop the SVM is its high training time and it is tricky to figure out the appropriate kernel[10].

Conclusion

To conclude, it can be inferred from the above case study that the type of data and characteristics of the features are pivotal in deciding the technique that needs to be implemented. If the data has linear variables and less number of features, it is always convenient and accurate to rely on the logistic regression for cases such as ICU predictive modeling. In more complex cases, machine learning tools can be used. SVM, ANN (Artificial Neural Network, and techniques like fuzzy matching are often more accurate but are treated as "Black Box Models" and hence are less relied on). It has been generally observed that the high-bias classifiers are accurate for small dataset and low-bias are best for big dataset.

The case study also highlights one of the many applications of predictive analytics for targeting algorithms and how directly it can affect the revenue cycle and performance for the healthcare-provider. Greater accuracy in targeting resulting from insights generated by predictive models help providers in reducing the sunk cost due to targeting wrong population who were not a part of actual target list, or even targeting the population that were already existing and would have been in the system if they were not targeted. Secondly, it also reduces the chances to miss the potential patients such as in cases where potential patients were missed out due to inaccuracy of the targeting model. The third advantage is channel optimization, as predictive models even help in optimizing the channel cost, particularly for the expensive channels like phone calls or direct mail.

However, the scope of these predictive algorithms is not limited to only targeting, but can be leveraged in prediction of various diseases, medical adherence and appointment rescheduling algorithm based on the commonly available features in the hospital database.

EXL look deeper.®

References

[1] Four Use Cases for Healthcare Predictive Analytics, Big Data
by Jennifer Bresnick (http://www.predictiveanalyticsworld.com/patimes/four-use-cases-for-healthcare-predictive-analytics-big-data-0422153/5218/)

[2] Comparison of target selection methods in direct marketing.
by Sara Madeira. João M. Sousa. Technical University of Lisbon, Instituto Superior Técnico. Dept.

[3] Medical/Health Predictive Analytics – Logistic Regression
Clive jones- http://businessforecastblog.com/medicalhealth-predictive-analytics-logistic-regression/

[4] A Multivariate Logistic Regression Equation to Screen for Diabetes by Bahman P. Tabaei, MPH1 and William H. Herman, MD, MPH12

[5] Decision trees: an overview and their use in medicine by Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman University of Maribor – FERI

[6] The clinical decision analysis using decision tree by Jong-Myon Bae

[7] Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients
by Youn-Jung Son, RN, PhD, 1 Hong-Gee Kim, PhD, 2 Eung-Hee Kim, ME, 2 Sangsup Choi, PhD, 2 and Soo-Kyoung Lee, RN, Doctoral Candidate

[8] Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes
by Wei Yu, Tiebin Liu, Rodolfo Valdez1, Marta Gwinn and Muin J Khoury

[9] Comparison between SVM and Logistic Regression: Which One is better to Discriminate?
by Diego Alejandro Salazar, Jorge Iván Vélez, Juan Carlos Salazar

[10] Comparison Logistic Regression vs. Decision Trees vs. SVM (http://www.edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part1/)

Acknowledgments

Radhika Sharma, Manager and Dr. Nakul Makkar, Project Manager

EXL (NASDAQ: EXLS) is a leading operations management and analytics company that helps businesses enhance growth and profitability in the face of relentless competition and continuous disruption. Using our proprietary, award-winning Business *EXLerator* Framework®, which integrates analytics, automation, benchmarking, BPO, consulting, industry best practices and technology platforms, we look deeper to help companies improve global operations, enhance data-driven insights, increase customer satisfaction, and manage risk and compliance. EXL serves the insurance, healthcare, banking and financial services, utilities, travel, transportation and logistics industries. Headquartered in New York, EXL has more than 24,000 professionals in locations throughout the United States, Europe, Asia, Latin America, Australia and South Africa.

Email us: lookdeeper@exlservice.com On the web: EXLservice.com

GLOBAL HEADQUARTERS
280 Park Avenue, 38th Floor, New York, NY 10017
T: +1.212.277.7100 • F: +1.212.277.7111

United States • United Kingdom • Czech Republic • Romania • Bulgaria • India • Philippines • Colombia • South Africa