Proceedings of the 2015 INFORMS Workshop on Data Mining and Analytics
M. G. Baydogan, S. Huang, A. Oztekin, eds.

MERGING BUSINESS KPIs WITH PREDICTIVE MODEL KPIs FOR BINARY CLASSIFICATION MODEL SELECTION

Matthew A. Lanham & **Ralph D. Badinelli**
Virginia Polytechnic Institute and State University
Department of Business Information Technology (0235)
1007 Pamplin Hall, Blacksburg, VA 24061
lanham@vt.edu/ralphb@vt.edu

Abstract

This study provides an example from a national retailer using binary classification techniques to model the propensity that a product will sell within a certain time horizon. We posit that a firm performing predictive analytics should consider the statistical performance, as well as the performance that a set of potential models will have with respect to business indicator(s) that the model is supporting. Model assessment statistics (e.g. AUC, overall accuracy, etc.) are important metrics that gauge how well a model will predict future observations, but we have discovered that using them in isolation is insufficient when deciding which model performs optimally with regard to the business. Modeling the propensity that a product will sell in a particular store using several binary classification techniques, we capture their traditional assessment statistics and point out which model would likely be chosen. We show that a better solution would be to build a decision model that selects the best forecasts using both traditional assessment statistics and business performance.

Keywords: Analytics, Model Assessment, Model Selection

Introduction

Model assessment statistics (e.g. AUC, overall accuracy, etc.) are important and commonly used to gauge how well a model will predict future observations. We posit that using these statistics in isolation are insufficient when deciding which model performs the best with regard to the actual business problem.

The analytics movement continues to gain traction among firm executives, so it is more important than ever that practitioners are linking the complicated algorithms they are using to actual business outcomes those algorithms are supporting. Whether practitioners are doing this correctly or not is unclear, but what is becoming apparent is that executives are looking at the results of previously made decisions and their corresponding results from their Business Intelligence (BI) platforms.

BI is an umbrella term to describe analytical concepts and methods to improve managerial decision-making by using fact-based support and reporting systems [1]. Recently the focus has shifted slightly from BI to Business Analytics (BA). BA and BI are often used interchangeably, but BA is really a component of BI that provides the value from data analyses and modeling techniques [2]. BA is "*the scientific **process** of transforming data into insight for making better decisions [3]*." The effective use of BI reporting systems by upper management is providing them insights in regards to the BA solutions being provided, and used by, decision-makers further down the decision-making hierarchy. In turn, BI is providing a useful feedback loop to the BA practitioners and decision-support solutions they are developing.

We structure this paper by turning the business problem into an analytics problem, describe the model methodologies we employed, detail our decision model selection procedure, and provide some results. Lastly, we discuss the positive impacts and insights of our solution and how we are working to validate our decision model using different constraints and parameters.

Business Problem to Analytics Problem

A retailer's assortment decision asks what are the optimal products to offer in a particular location, how much inventory to carry for those products [4]. The process of determining an assortment plan can happen at various points in a year, the assortment decision involves listing and delisting products over time as consumer demand changes [5]. Thus, decision-makers (e.g. category managers) require having parameters that gauge a products future selling propensity to help support their assortment decision effectively.

In cases such as this, refining the business requirements with decision-makers can lead to important discoveries that lead to better decision-support. We discovered that our predictive forecasts require certain unique characteristics to be used effectively. First, it is important to the assortment decision that the probabilities are discriminatory in nature. We do not mean discriminatory in the classical binary classification sense (e.g. True-positives/True-negatives), but rather the probabilities are spread over the entire $[0,1]$ probability space. The reason for this is to allow the decision-maker to be able to identify one SKU that is better than a competing substitutable SKU. For example, if a grocery store had three ketchup products to choose from to put into a store (e.g. Heinz, Hunts, Store brand), but had a constraint to put only two into the assortment, having probabilities that are the same value present discriminatory issues.

The second requirement we discovered as we were formulating potential analytical solutions to this problem was that the category mangers were not the only stakeholder using such measures for decision-making purposes. It turned out that such "propensity-to-purchase" aggregated forecasts are used by executives in the decision-making hierarchy to support strategic planning initiatives (e.g. grow category A by 5%). These executives were not only using these measures, but were able to identify from their BI suites which probability forecasts were performing as expected and which were inconsistent with actual business performance (e.g. percentage sold).

Methodology Selection

Since our response is categorical (e.g. sell or not sell), the analytical problem is deemed a binary classification problem under the predictive modeling umbrella. When binary classification algorithms are employed they try to classify an object (e.g. SKU) into one of only two possible groups. One interesting aspect is the decision cut-off need not be fifty percent, which is common practice. Any threshold could be employed based on a variety of reasons (e.g. unbalanced data set, problem-specific, etc.). The idea is that the estimated probability value greater than the specified threshold results in a product being classified as a seller, while less than the threshold a non-seller.

To generate our probabilities we chose several commonly used binary classification algorithms, such as logistic regression, classification tree, C5.0 decision tree, Quest decision tree, CHAID, decision list, linear discriminant model (LDA), artificial neural networks using a multi-layer perceptron and a radial basis function, as well as a support-vector-machine (SVM) using a radius basis function. We also trained boosted and bagged versions of these models, as well as a heterogeneous ensemble model that used weighted predictions from the other models based on their corresponding model confidence/accuracy.

Data & Modeling Building

The data set investigated consists of 49,656 records entailing one product category from a national retailer. Each observation contained a store-SKU combination. For example, store i has j records, where each record is a unique SKU that has sold or not sold over some time horizon. The attributes used or time-horizon employed could not be disclosed due to confidentiality concerns, but the attributes entail measures with respect to the store, SKU, and demographic profiles. All records were products that were stocked and sold, stocked and did not sell, or were purchased in some other fashion (e.g. online, in-store, telephone) and delivered to a store for customer pickup in a store. The general predictive model employed by all algorithms used was as follows:

$$Y_{ij} = f(attribute\ list), \text{ where } Y_{ij} = \begin{cases} 1\ if\ sold\ one\ or\ more\ units\ of\ sku\ i\ in\ store\ j \\ 0\ if\ did\ not\ sell\ one\ or\ more\ of\ sku\ i\ in\ store\ j \end{cases}$$

All algorithms were trained using a 50/50 balanced training data set due to it being slightly unbalanced. Balancing is common practice when data sets are unbalanced as the algorithms will tend to build a model classifying the majority class better than the minority class. The data was trained and assessed using a 70/30 percent training/testing partition with 10-fold cross-validation.

To assess the models we report common statistical performance measures, such as area under the curve (AUC), overall classification accuracy (using a 50% cutoff), as well as lift and profit measures. In practice many competing models are generated and the retailer will often choose one among the set.

Results

Plotting the percentage of products sold versus the forecasted probabilities binned together in five percent bins reveals the actual business performance to expected business performance over that specific time horizon. Ideally if the models are producing correct probabilities over this horizon they would follow a 45 degree line as shown via the black dotted line in the Figure **1** plots below. The Linear Discriminant Analysis (LDA) model is very consistent for each probability bin, but consistently underperforms compared to the actual percentage that sold. The boosted neural network radial basis function (BANN_RBF) model revealed approximately 35 percent of SKUs within every probability bin previously sold. Since the number of SKUs can vary within each probability bin, such an occurrence can happen, but may not be obvious based on the plots alone.

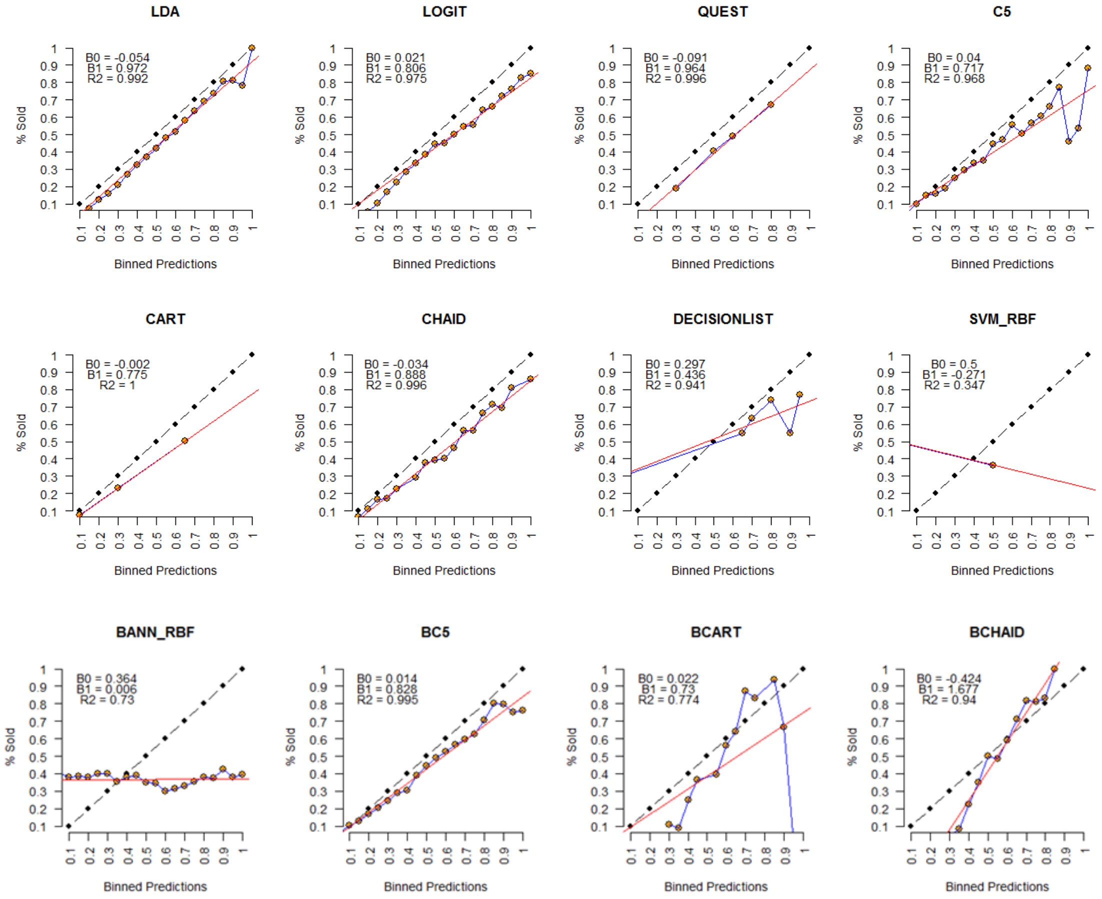

Figure 1: Percentage sold of products versus estimated probability to sell for each model (Note that not all models are shown due to proceedings length restrictions).

Interestingly, most of the models perform well statistically (e.g. ROC/AUC, etc.) as shown in Table **1**, but some perform better than others with regard to actual business performance.

	Traditional Assessment & Model Selection Measures					Customized Business Assessment & Selection Measures		
Model	Max Profit	Max Profit Occurs in (%)	Lift	Overall Accuracy	Area Under Curve	B0-Intercept	B1-Slope	R-squared
Logit	34945 / 6550	54 / 46	1.58 / 1.69	70.22 / 71.15	0.785 / 0.789	0.021	0.806	0.975
CART	31526 / 4582	69 / 67	1.26 / 1.31	68.38 / 66.65	0.693 / 0.696	-0.002	0.775	1.000
C5.0	40877 / 6749	51 / 43	1.62 / 1.67	73.94 / 71.48	0.815 / 0.793	0.040	0.717	0.968
Quest	31337 / 5700	57 / 45	1.49 / 1.58	68.28 / 67.98	0.722 / 0.731	-0.091	0.964	0.996
CHAID	34051 / 6009	56 / 41	1.54 / 1.64	69.87 / 68.52	0.777 / 0.772	-0.034	0.888	0.996
Decision list	28533 / 5583	30 / 28	1.55 / 1.61	66.66 / 69.12	0.679 / 0.68	0.297	0.436	0.941
LDA	33005 / 6130	52 / 40	1.54 / 1.63	69.14 / 69.75	0.767 / 0.769	-0.054	0.972	0.992
ANN (MLP)	35275 / 6695	50 / 43	1.59 / 1.7	70.57 / 71.21	0.789 / 0.795	-0.018	0.857	0.999
ANN (RBF)	29055 / 5150	56 / 42	1.46 / 1.56	66.82 / 66.72	0.733 / 0.74	-0.012	0.850	0.993
SVM (RBF)	1340 / -8	100 / 1	1 / 1	49.6 / 55.22	0.501 / 0.501	0.500	-0.271	0.347
Boosted CART (30)	35387 / 6395	51 / 36	1.58 / 1.68	70.59 / 70.01	0.784 / 0.782	0.022	0.730	0.774
Boosted C5.0 (30)	7330 / 7330	42 / 42	1.75 / 1.75	73.06 / 73.06	0.818 / 0.818	0.014	0.828	0.995
Boosted Quest (30)	35892 / 6830	50 / 44	1.6 / 1.7	71 / 71.21	0.777 / 0.781	-0.169	1.146	0.945
Boosted CHAID (30)	41375 / 6455	51 / 47	1.66 / 1.69	74.29 / 71	0.821 / 0.795	-0.424	1.677	0.940
Boosted ANN MLP (10)	37210 / 6805	52 / 45	1.61 / 1.7	71.78 / 71.57	0.802 / 0.801	-0.015	0.869	0.998
Boosted ANN RBF (10)	32210 / 5885	52 / 45	1.52 / 1.6	68.66 / 69.76	0.757 / 0.763	0.364	0.006	0.730
Bagged CART (30)	31851 / 4933	66 / 63	1.28 / 1.33	68.15 / 66.22	0.701 / 0.705	0.042	0.618	0.968
Bagged Quest (30)	31110 / 5701	58 / 44	1.48 / 1.58	68.21 / 67.98	0.721 / 0.732	-0.096	0.974	0.996
Bagged CHAID (30)	35525 / 6285	54 / 45	1.56 / 1.64	70.83 / 69.85	0.785 / 0.782	-0.064	0.965	0.992
Bagged ANN MLP (10)	38000 / 7105	48 / 37	1.64 / 1.74	72.21 / 72.22	0.81 / 0.81	-0.010	0.852	0.993
Bagged ANN RBF (10)	32210 / 5885	52 / 45	1.52 / 1.6	68.66 / 69.76	0.757 / 0.763	0.364	0.006	0.730
Heterogeneous Ensemble	-	-	-	73.06 / 73.06	0.74 / 0.74	0.601	-0.471	0.646

Table 1: Statistical and business assessment and selection measures

In practice, often the model having the best predictive model assessment statistic would be chosen. In this case, the Boosted C5.0 decision tree is best (i.e. AUC = 81.8%). From a business perspective, this model performs well as can be seen above. The values follow a linear trend line (i.e. R-squared = 0.995) and also has probabilities nicely dispersed across the entire probability space. However, the probabilities generated are pessimistic (i.e. slope = 0.828) compared to actual sales performance. Next, we show how we solve this problem to create a more robust solution both with regard to the business and even leads to increased overall classification accuracy.

Deployment - Decision Model

Our solution takes advantage of all of the intelligent models available. We accomplish this by selecting the model having the closest probabilities to the actual percentage sold for each five percent bin. We incorporate constraints so that poor models (i.e. do not generate probabilities in at least half the bins, have an AUC less than 60%, etc.) are excluded from being used even if they have probabilities that are perfectly aligned with actual business performance.

<u>Terms and definitions</u>:
x_{ij} = the probabilities from model i are chosen for bin j, $x_{ij} \in \{0,1\};\ i = 1,..,N;\ j = 1,..,M$
$\beta_{0,i}$ = the estimated intercept parameter corresponding to liner regression model $i;\ i = 1,..,N$
$\beta_{1,i}$ = the estimated slope parameter corresponding to linear regression model $i;\ i = 1,..,N$
ρ_i = the r-squared statistics corresponding to liner regression model $i;\ i = 1,..,N$
α_i = the area under of the curve for model i based on the testing set; $i = 1,..,N$
τ_i = the overall training accuracy of model i based on a 50% decision cutoff threshold; ; $i = 1,..,N$
φ_i = the out of sample testing accuracy of model $i;\ i = 1,..,N$
y_{ij} = the 45 degree line value for model i and bin $j;\ i = 1,..,N;\ j = 1,..,M$
$\hat{y}_{ij}$ = the percentage sold for model i within each bin $j;\ i = 1,..,N;\ j = 1,..,M$
K = large value to penalize the objective function
θ_{ij} = squared error of % sold to 45 degree line for model i for bin j, such that = $\begin{cases} (\hat{y}_{ij} - y_{ij})^2\ if\ \hat{y}_{ij}\ exists \\ K\ otherwise \end{cases}$

<u>Objective function</u>:

$$min \sum_j \sum_i \theta_{ij} * x_{ij}$$

<u>Constraints</u>:

(1) $x_{ij}(\varphi_i - \tau_i) \leq 0.05 \quad \forall\, j, i$ (use only valid models)
(2) $x_{ij}(\tau_i - \varphi_i) \leq 0.10 \quad \forall\, j, i$ (ignore overfit models)
(3) $-0.45 \leq x_{ij}\beta_{0,i} \leq 0.45 \quad \forall\, j, i$ (reasonable intercept)
(4) $0.70 \leq x_{ij}\beta_{1,i} \leq 1.30 \quad \forall\, j, i$ (reasonable slope)
(5) $0.50 \leq x_{ij}\rho_i \leq 1\ \forall\, j, i$ (linear fit model points close to line)
(6) $x_{ij}\alpha_i \geq 0.60\ \forall\, j, i$ (probabilities can only come from intelligent models)
(7) $\frac{M - \sum_j \delta(\theta_{ij} - K)}{M} \geq 0.5$ (only use models having probabilities that exist in more than half the $[0,1]$ space)
where $\delta(\theta_{ij} - K) = \begin{cases} 1\ if\ \theta_{ij} - K = 0 \\ 0\ if\ \theta_{ij} - K \neq 0 \end{cases}$
(8) $x_{ij} \in \{0,1\}$ (binary decision)

Our decision model led to using forecasts from ten different predictive models. Among the 22 models we generated, nine were excluded from consideration because they did not meet the inclusion criteria (e.g. constraint (6)). Models excluded came from the basic set, boosted set, bagged set, as well as the heterogeneous ensemble model. The final set of models selected for each bin as shown in Figure **2** achieved a classification accuracy of 76.6%, which is 3.54% better than the best model, but more importantly leads to propensity to purchase estimates that follow more closely with business performance (e.g. R-square of 99.1%).

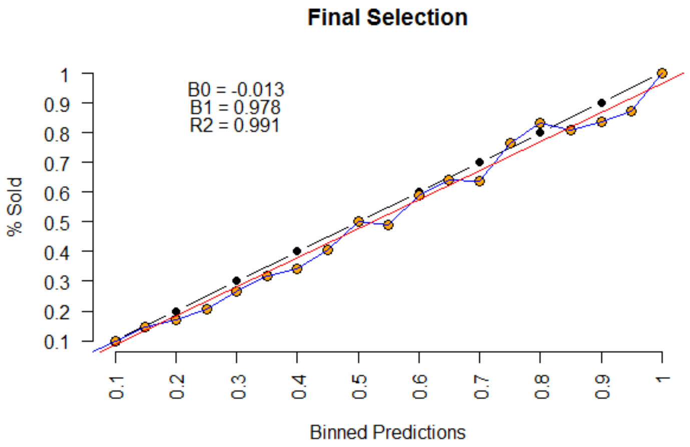

Figure 2 Final Selection Results

Conclusions & Future Research

Modeling the propensity that a product will sell in a particular store using several binary classification techniques and reviewing their traditional assessment statistics we identify the model that would have likely be chosen and put into production in practice. Interestingly, the optimal model does not perform optimally with regard to the business. If many competing models are generated, which is common, a retailer could choose to strategically use all intelligent forecasts that are most likely to match business performance. This can not only reduce uncertainty for planning purposes, but also helps achieve buy-in with direct decision-makers that rely on the decision-support solutions the analytics professionals are generating.

We are currently working on validating different decision models having different constraints and parameters to identify a model that performs optimally for many product categories.

References

[1] Chen, H., R.H. Chiang, and V.C. Storey, *Business intelligence research.* MISQ Special Issue (forthcoming), 2011.

[2] Wixom, B., et al., *The current state of business intelligence in academia.* Communications of the Association for Information Systems, 2011. **29**(1): p. 16.

[3] INFORMS, *INFORMS Certified Analytics Professional (CAP) Examination Study Guide*. 2014: www.informs.org.

[4] Kök, A.G., M.L. Fisher, and R. Vaidyanathan, *Assortment planning: Review of literature and industry practice,* in *Retail Supply Chain Management*. 2015, Springer. p. 175-236.

[5] Hübner, A.H. and H. Kuhn, *Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management.* Omega, 2012. **40**(2): p. 199-209.