EVALUATING STOCHASTIC COST-BENEFIT CLASSIFICATION MEASURES FOR A RETAILER'S ASSORTMENT MIX DECISION

(Submitted to the Graduate Student Paper Competition – PhD Track)

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

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

ABSTRACT

In this paper we show the expected gains that a retailer might achieve by using stochastic cost-benefit analysis to select the products incorporated into their assortment planning decision as compared to traditional, statistical performance measures for model assessment in isolation. The motivation for this research is that the assortment decision is considered one of the most important decisions that a retailer will make due to its direct link to sales, inventory costs, margin, etc. We investigate naïve binary classification models trained using different rebalance techniques to identify product "sellers" from "non-sellers." This approach can be used to rank products based on raw propensity scores or be used to reduce the evaluation space when faced with having to model hundreds of thousands of products and estimating their substitution behavior. Using product category data from a national retailer, we train and evaluate different rebalancing-model sets and compare the outcomes from a statistical perspective, in addition to a financial perspective that incorporates the retailer's revenue and costs for the products placed in the assortment.

Keywords: Assortment Planning, Binary Classification, Model Validation, Cost-Benefit Analysis

INTRODUCTION

An assortment is the product set carried in each store at any point in time [1]. This geospatially-customized set will require modification over time due to changes in consumer preferences. The domain of assortment planning has traditionally lied in the strategic marketing planning domain, whose goals are geared toward product categories or supplier contracts, which make for longer planning horizons [2]. The primary task involved in assortment planning includes listing and delisting products so that an optimal assortment remains over time.

Assortment planning research focus has changed over time as the area has evolved and new approaches to modeling consumer demand are examined. Historically, assortment planning has been a process employed to find the optimal set of products to carry and amount of inventory to maintain of each product [3], but today it is viewed more specifically as making product decisions based on consumer choice behavior and substitution effects [2]. The research in this area continues to grow because the retailers' assortment mix is and will always be one of the most important decisions faced by retailers. This is because of the impact that the set of products carried and not carried can have on key business performance indicators (KPIs), such as overall sales, inventory costs, margin, etc. [1, 4]. Moreover, retailers with the ability to modify assortment decisions as demand evolves can create competitive advantages, as well as differentiate themselves from their competitors [5].

From a decision modeling perspective, the objective of any assortment planning model is to determine which products to carry in a location (e.g. store, HUB, DC) in order to maximize sales. In addition to this performance measure, various constraints must be satisfied. Budgetary constraints are imposed on each category manager (i.e. the assortment decision-maker) that limits what products they can add during their seasonal line review. This budget can be decomposed into the number of new[1] products they can add to their categories' assortment during a line review and the products already stocked in each store that will remain in their assortment until they are sold, marked down, moved to another location, or returned to the distribution center.

Shelf space is another constraint that limits the number (and amount) of product that can be stocked on the shelf per category. One could easily classify the assortment decision as a classical knapsack of knapsacks OR problem with the physical space of the store and respective category shelf space within the store being the primary physical constraints of the optimization problem. Due to the capacity constraints, a retailer will not have to ability to stock in every location every potential product a consumer may desire [4]. Even if it was physically possible, it would not be financially feasible. Strategic constraints are also often incorporated into the assortment decision that account for competition and supply chain resiliency. Examples include having more product coverage for stores having more local competition and having multiple vendors for each product type. Recent research in this area has shown that if a rival store exists within its geographical market, the presence of the rival can have an effect on product variety and product overlap. Specifically collocated rivals (less than a mile apart) are more apt to differentiate their assortments by having less overlapping products than their distant rivals (a competitor further than a mile away, but within 10 miles) [6].

The most challenging aspect of assortment planning is being able to solve the decision model because the combination of potential assortments is large given the vast number of stock-keeping-units (SKUs), which leads to a problem having NP-Hard computational complexity [7]. Aside from solving the decision model, we find a more interesting challenge is to identify consumer demand preferences among the vast set of SKUs. Tackling this challenge gets into the hot research area of Big Data Analytics (BDA).

Big data is the massive amounts of data that firms collect via customer databases, web crawlers, server logs, social media, and other devices [8]. The big data problem can be divided into three basic components known as the 3Vs: volume, velocity, and variety. Retailers today are not only capturing more internal measures (e.g. web clicks), but are capturing vast external sources of information coming from new sources such as social media (e.g., Facebook, Twitter, etc.), machine sensors, weather forecasts, and GPS location tracking making data grow faster than ever before. The total amount of potentially insightful data nearly doubles every two years, and thus has been referred to by some as "*Moore's Law of Marketing* [9]." The largest retailer in the world being Walmart collects approximately 2.5 petabytes of customer transaction data every hour [10]. The velocity/speed of all this incoming data makes data processing expensive to employ using traditional in-house hardware and software analytics tools that must process the data in a timely fashion [11]. Lastly, the variety of the data, most notably unstructured data, leads to issues of storing and analyzing it, compared to structured data stored in relational databases. This has led to new technologies, frameworks, and languages to be developed, such as Hadoop, Map Reduce, and Pig to name a few.

According to SAS, "*Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions* [12]." Due to the interest and potential of BDA, the Data Scientist role has emerged to tackle the challenge of finding patterns in diverse data sources with the goal of better understanding consumer behavior and to identify

[1] New products are defined here as either 1) completely new products purchased from a vendor or 2) products that are "new" to the a particular store because they have never been stocked in that particular location

new opportunities [13]. We posit that collecting data is the easy part, but identifying precise and creative approaches to analyze the data collected is the true challenge. As BDA become more mainstream, it will change established views on the value of experience, expertise, and management practice [10]. Fortunately, some firms are not afraid to modify their organizational culture to a more analytics-rich culture.

According to Andrew McAfee and Erik Brynjolfsson of MIT, *companies that inject big data and analytics into their operations show productivity rates and profitability that are 5% to 6% higher than those of their peers* [14]." Thomas Davenport found that among the 50 firms he interviewed that were experimenting with BDA are achieving results in the form of reduce costs, faster decisions, and new products and services [15]. We expect the gains from implementing the solution we propose in this study can lead to reduced costs, and possibly faster decisions.

With regard to employing BDA to improve assortment planning, we have found nothing in the literature. The majority of papers focus on toy problems having assortment planning test cases averaging less than or equal to 29 items which are unrealistic in practice [2]. One of the fundamental requirements needed to support the assortment decision are accurate measures of demand, which may be unit forecasts, propensity to sell more than some specified number of units in a particular store, as well as consumer choice substitution behavior. Unit forecasts are typically univariate or multivariate time series models such an error-trend-seasonality (ETS) models, or advanced ARIMIA models. Propensity to sell models are binary classification approaches typically used to rank products on a [0,1] standardized scale using binary classification techniques such as logistic regression, decision trees, etc. The binary response is calculated by specifying winners, number of previously sold units greater than x as one class (i.e. "sellers"), and the other observations less than x as a different class (i.e. "non-sellers"). Consumer-choice substitution behavior modeling has been popular as of late in the assortment planning literature, where methodologies such as multinomial logistic regression, location-choice, and exogenous demand are often used to estimate the probability that one product in a set of substitutable products will be chosen where the probabilities for each substitutable set equal one. However, binary classification models could be used to rank an entire category of products based on their propensity to sell, or to reduce the modeling set to something more reasonable when identifying potential substitutes using multi-classification techniques (e.g. multinomial logistic regression). Based on the previously mentioned potential benefits of BDA and the need for better demand predictive models to support the assortment decision, we believe this paper provides a basic, yet novel, approach to assessing the performance of the predictive models based on financial measures.

We structure this paper by providing a brief synopsis of the academic literature, describe the research design methodology we employ, discuss our results, and lastly discuss our conclusions, practical limitations and research we are currently performing to make our proposed solution extendable to other modeling endeavors where financial performance considerations must be considered.

LITERATURE REVIEW

With regard to employing BDA to improve the assortment planning, we have found nothing in the literature. According to Kök, Fisher [1] there still lacks a dominant solution for assortment planning, as the area is still relatively new, which is providing "*a wonderful opportunity for academia to contribute to enhancing retail practice*." Academic research on improving the assortment decision has focused on formulating an optimization model that selects the best sets of products to carry and their respective inventory levels. Due to the computational complexity of the number of products to choose from, a single category or subset of similar products is usually investigated. Research has mostly focused on one assortment for a retailer, even though retailers will regularly have different assortments for different stores due to the differences in customer preferences. According to Hubner and Kuhn, shelf space

planning research has plateaued over the past three decades, but assortment planning continues to gain interest in the operations research field because of a partial incorporation of substitution effects in their model [2].

From an operational perspective, retail product and shelf allocation are made with little regard to a company's overall strategies or cross-functional effects [16]. We have found such decisions are usually made in their respective departmental silo and then passed on as parameters to assortment planners working to derive a customized assortment per store, as well as inventory management which is tasked with getting the product from suppliers into the DCs, HUBs, and stores.

From a Decision Support System (DSS) perspective, there is a great need because of the challenge of modeling descriptive, predictive, and prescriptive components required of the assortment decision cohesively at such a scale. Moreover, the competitive nature of business continues to intensify as consumers become more research savvy leading retailers to focus more on their customer's needs to provide them the products and services they desire, while still achieving operational competence. It has been shown that assortment and shelf space planning models are not comprehensively united into commercial software, and the software vendor packages mainly focus on large-scale data processing and less on intelligent decision making algorithms [2]. In the end, retailers are seeking more robust tools that leverage information technologies to provide them scalable platforms to help them model product demand more efficiently and make optimal assortment modifications as demand changes. These tools must not only consider the technical, statistical, and computational aspects of the modeling endeavors, but the business performance measures as well.

To figure out what to stock in a category, the retailer must have a firm understanding of what they want. The "*first moment of truth*," is what Procter & Gamble refer to as the moment in time after a shopper has arrived at a retailer's shelf [17]. At this point the retailer will either have the product that the customer desires, have a product substitute, or the customer will leave without making a purchase.

Most of the academic research that models purchasing behavior has focused on substitution behavior. Substitution is when a customer seeks a particular product at some venue, the product is not available, (because it is not carried or out of stock) and the customer then decides to purchase a similar product in its place. Being able to quantify substitution behavior is important when customers have a higher propensity to substitute within a category because the retailer need not have as much depth, nor a high in-stock service rate. However, when customers are less likely to substitute, more depth and high in-stock service levels are important to reduce the impact of lost sales [1].

There are three types of substitution: stock-out based, assortment-based, and utility-based substitution. Stock-out based substitution is when a customer shops recurrently for a product that has been purchased before (e.g. a daily consumable such as milk), but finds it out of stock so they purchase a different product. Assortment-based substitution is when a customer has identified a particular target product, possibly from what they observed as being offered in other stores or from advertisements. However, they cannot purchase it within a store because it is not carried so they purchase a different product. Stock-out and assortment-based substitution occurs frequently when consumers are shopping for daily consumables such as grocery products. Utility-based substitution is more of a speculative idea to understanding substitution behavior that claims that a consumer will purchase the product that yields them the highest utility if greater than the no purchase option is among a set of products available on the shelf. It could be that their target item is not on the shelf or there are other items that would yield them higher utility that they are unaware of, but since these products were not in the assortment or out-of-stock their purchase decision involved substitution effects. Consumers regularly make utility-based substitution purchasing decisions when shopping for clothing, consumer electronics, or auto parts. Products in an assortment may serve as substitutes, so the customer may purchase some product just to satisfy their needs and not their

actual wants. This leads to a condition where the demand for each product is influenced by the assortment of products that are offered [18]. The marketing literature has shown that when a consumer's target item is not available, anywhere between 45% and 84% of demand can be satisfied by substitution goods [19-22]. A consumer's particular characteristics, the situation, and the product itself will influence the expected substitution potential [22, 23].

An assortment should have the right depth (i.e. product categories) versus diversity (i.e. number of options within a category). However, being able to generate such an assortment is a challenging task. High selling items could be due to highly-correlated compliment or substitutable products that would not sell alone. There are other types of products that sell better when pooled with other products (e.g. chocolate syrup in the ice cream isle). Amazon.com is one example of an online retailer that has figured out how to do this well. Their recommended items section supported by their proprietary recommendation engine provides customers suggested items they are likely to want based on their viewing and purchasing behavior. These reasons alone show that demand forecasts must look beyond traditional approaches to modeling consumer demand (e.g. MNL), or simple time-series (e.g. ETS, ANOVA) model, and be savvy enough to identify opportunities to improve the assortment. For these reasons, and an acknowledgement that understanding substitution behavior is important, we venture other approaches to modeling product propensity and try to tie their traditional assessment measures back to the key business measures ($).

Binary Classification

In this study, we investigate binary classification predictive models. Binary classification algorithms fit a model having a binary response (e.g. No/Yes, 0/1, etc.). When the model is a fit, a continuous prediction for each record is generated that lies between 0 and 1. These predictions are then compared to a specified decision cutoff criterion, which allows each prediction to be classified as a member of only one of two possible disjoint classes. For example, a predictive value of 0.72, with a specified decision cutoff criterion of 0.60 would classify this observation to "Class 1", instead of "Class 0."

There are many binary classification algorithms coming from the statistics and computer science/machine learning literatures, and each has its own optimization performance measure that it is trying to minimize. Each model can perform better than another model on different datasets due to characteristics of the predictors, the dimensionality of the data, and how balanced the classes are when training a model.

Our binary classification models are employed on previously stocked stock-keeping-unit (SKU) that have sold over a particular time interval. Each SKU that has sold over a specified number of units as deemed by the retailer are considered "sellers" (or Class 1), while those that have not sold more than a specified number of units over a certain time interval are labeled as "non-sellers" (Class 0).

Class Imbalance

Class imbalance occurs when one class has more observations than another class. For example, training a model with 100 records where 70 had a response value of "Class 1", while the remaining records had a response value of "Class 0", which would yield a minority class imbalance of 30%. Studies have shown that not taking into account the class imbalance can have negative consequences on model estimation [24]. Essentially, the imbalance makes it difficult for the binary classification to learn, which usually leads to correctly classifying majority class records and not the minority class records [25]. Today, there is not one technique that works optimally well for all data sets and researchers continue to work on creating a unified rebalancing framework to tackle this problem [26]. He and Garcia [27] provide a throughout review of this area. In our study, we investigate non-balancing, down-Sampling, up-Sampling, SMOTE, and ROSE.

Non-balancing means that after the entire data set is randomly partitioned into training and testing sets, the training set will usually have a class imbalance percentage that follows closely, if not exactly, to that of the overall data set. In this case, the training data set is used to train the binary classification model. Down-sampling is when a sample of the majority class is used so that the number of records in both classes is the same. Likewise, Up-sampling resamples the minority class until the numbers of records for both classes are the same. Up-sampling will always lead to more training records which can be a positive when training a model. However, when the data dimensionality is large it can be computationally time consuming and must be considered by the retailer so that predictions can be delivered to decision-makers as scheduled. Synthetic Minority Over-sampling Technique (SMOTE) over-samples the minority class by creating "synthetic" records rather than resampling the minority class with replacement. Finally, Random OverSampling Examples (ROSE) is another common rebalancing technique that employs a systematic framework for correcting learning issues that arise from unbalanced data by employing a smoothed bootstrap resampling methodology [26].

Model Assessment

When predictive models are constructed they must be empirically validated. Typically, the process involves fitting/training a model on one data set, and then once the model is constructed to feed in another data set that was not used to train the model commonly referred to as a testing set. This cross-validation procedure provides a proxy of truth about the validity of the model as well as gauges the expected performance of future predictions when new data are used.

Next, we will discuss traditional statistical measures of model performance, such as the confusion matrix and how such matrices lead to the construction of the Receiver Operating Characteristic (ROC) curve. The ROC curve is the most frequently used metric used to compare and assess different models. We then introduce stochastic cost-benefit analysis measures and provide a basic explanatory example before employing it on a national retailer's product category.

Confusion Matrix & Statistics

The confusion matrix, as shown in Figure 1, is a cross-tabulation table that provides a gauge of how well a model's predictions were classified compared to the response. Since the response only has two possible classes in our case (i.e. sold vs. not sold), our predictions, in the form of a probability between [0,1], are assigned to one of these potential classes. The assignment is based on the modeler's specified cutoff threshold, which is most often 0.5. This means that if a prediction realizes a value of 0.67, it will be assigned to the "seller" class or "Class 1" because the prediction is larger than the threshold parameter. Likewise, a value of 0.49 would be assigned to "Class 0" or "non-seller" and so on for all probability predictions.

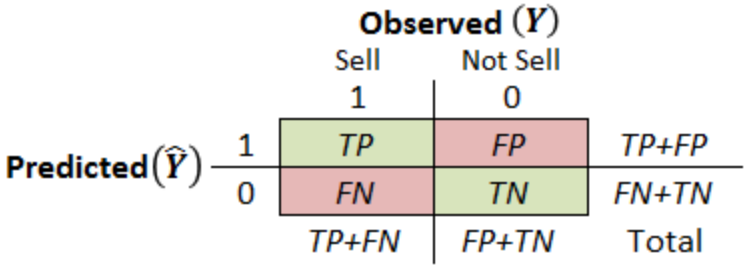

Figure 1: Confusion Matrix

A large number of true-positives (TP) and true-negatives (TN) on the diagonal of the matrix and a small number of false-positives (FP) and false-negatives (FN) provides an indication that the model used performs well at identifying SKUs that will sell or not sell in our context.

Many statistics can be calculated form this simple table. The overall accuracy of the model can be captured by $(TP + TN)/Total$ and is one benchmarking target often cited in practice to gauge performance. Interestingly, overall accuracy does not provide any distinction about the type of error the model is making and the prevalence that the SKU will sell. Some SKUs will inherently have higher or lower sales frequencies than other SKUs within their part type category. These differences in prevalence force us to measure many other important accuracy measures allowing us to make model improvements with the goal of continuous process improvement. For example, what is more expensive to a retailer's business – incorrectly putting a SKU in a store and it not selling or not putting a product in a store when it would sell? None of the statistical performance measures shown in Table 1 capture the financial business performance KPIs, such as expected profit of correctly classifying one SKU versus incorrectly classifying another SKU.

Statistic	Alternative Name	Description	Formula
Overall Accuracy		Probability that a SKU is classified as a seller and non-seller correctly	$\frac{TP + TN}{Total}$
Sensitivity	True-Positive Rate; Recall	Probability that a retailer will predict a sell when there is actually sell	$\frac{TP}{TP + FN}$
Specificity	True-Negative Rate	Probability that a retailer will not predict a sell when there is not a sell	$\frac{TN}{FP + TN}$
Type I error rate	False-Positive Rate; 1- Specificity	Probability that a retailer will predict a sell when there is not a sell	$\frac{FP}{FP + TN}$
Type II error rate	False-Negative Rate	Probability that a retailer will not predict a sell when there is actually sell	$\frac{FN}{TP + FN}$
Positive Predictive Value (PPV)	Precision	Probability that the SKU did sell when it actually sold (an unconditional analog to sensitivity, which takes into account the event's prevalence)	$\frac{TP}{TP + FP}$
Negative Predictive Value (NPV)		Probability that the SKU did not sell when it actually did not sell (an unconditional analog to specificity, which takes into account the event's prevalence)	$\frac{TN}{FN + TN}$
Positive Likelihood Ratio (PLR)		Ratio between the probability that the retailer predicts a sell when there is a sell and the probability that a SKU will sell given it actually did not sell	$\frac{Sensitivity}{1 - Specificity}$
Negative Likelihood Ratio (NLR)		Ratio between the probability that a retailer predicts a SKU will not sell given it actually sold and the probability that the retailer predicts that the SKU will not sell given that it did not sell	$\frac{1 - Sensitivity}{Specificity}$
Youden's J Index Youden [28]	J Index	The proportion of correctly predicted samples for both the seller and non-seller groups (an alternative to the ROC curve)	*Sensitivity + Specificity - 1*
Cohen's Kappa	Kappa	Historically been used to assess agreement among two raters but is appropriate to the retailer in this context as well [29]. If we let O = observed accuracy and E = expected accuracy, we can calculate Cohen's Kappa statistic based on the confusion matrices' marginal totals. Let, $O = TP + TN$ and $E = \frac{(TP+FP)*(TP+FN)}{(Total)} + \frac{(FN+TN)*(FP+TP)}{(Total)}$	$\frac{O - E}{Total - E}$

Table 1: Statistical measures to benchmark from our binary classification models

Area Under the Receive Operating Characteristics Curve

The Receiver Operating Characteristic (ROC) curve is created using two of the model assessment statistics in Table 1that were generated from the confusion matrix. The primary extension of the confusion matrix and corresponding statistics is that the ROC curve is based on calculating a confusion matrices based on different cutoff thresholds. As explained previously, the typical cutoff to assign probability predictions to classes is 0.50, but if one were to calculate and store confusion matrices having cutoffs ranging from 0 to 1 (e.g. 0.01, 0.02, …, 0.99) one could evaluate model classification performance for these different scenarios.

Having many confusion matrices with associated statistics could be difficult to analyze, thus the ROC plots sensitivity versus the type I error rate, also known as one minus specificity for each cutoff threshold as shown in Figure 2.

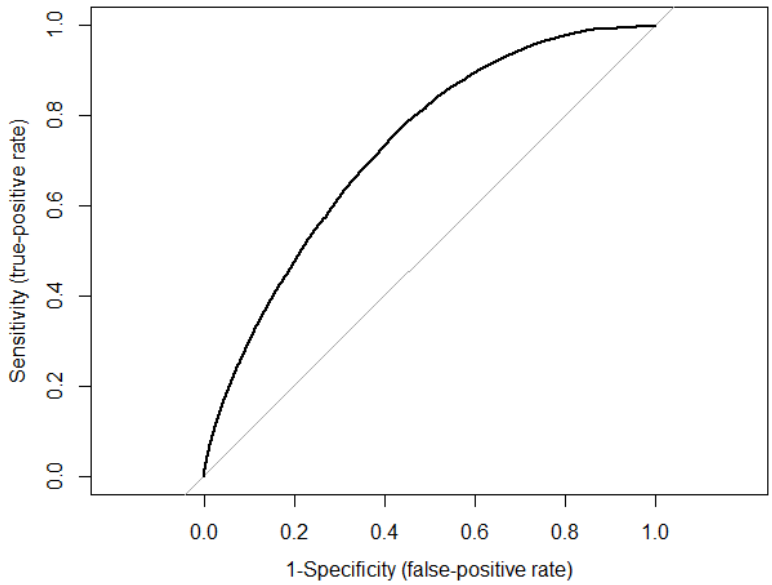

Figure 2: Receiver Operating Characteristic (ROC) curve

Figure 2 shows a nice smooth curve because the granularity of cutoff thresholds ranged from 0.001 to 0.999. In the context of our study, the sensitivity statistic estimates the probability that the retailer will predict a seller when the SKU was indeed actually a seller (i.e. sold), so it is a true-positive rate. A retailer would want sensitivity to be as high as possible, but the higher the sensitivity becomes leads to a negative tradeoff with the false-positive rate, meaning the retailer is more likely to predict and classify a SKU as a seller when in fact it is a non-seller. This is important to the retailer because stocking products that will not sell lead to additional inventory costs and take up shelf space for products that could sell. Moreover, the longer seasonal effects of having to markdown unwanted products hurts the retailer's margin, which based on the quantity of the markdowns, can lead to dramatic consequences to the retailer's bottom line.

To assess the overall performance of the model based on the ROC curve, the physical area under the curve (AUC) can be calculated. A curve having no classification ability will lie perfectly on the 45 degree line, thus the AUC is 0.50. Ideally, the retailer would prefer an ROC curve that goes from point (0,0) to (0,1) to (1,1) which would indicate perfect classification for any cutoff threshold and lead to an AUC of 1 as shown in Figure 3.

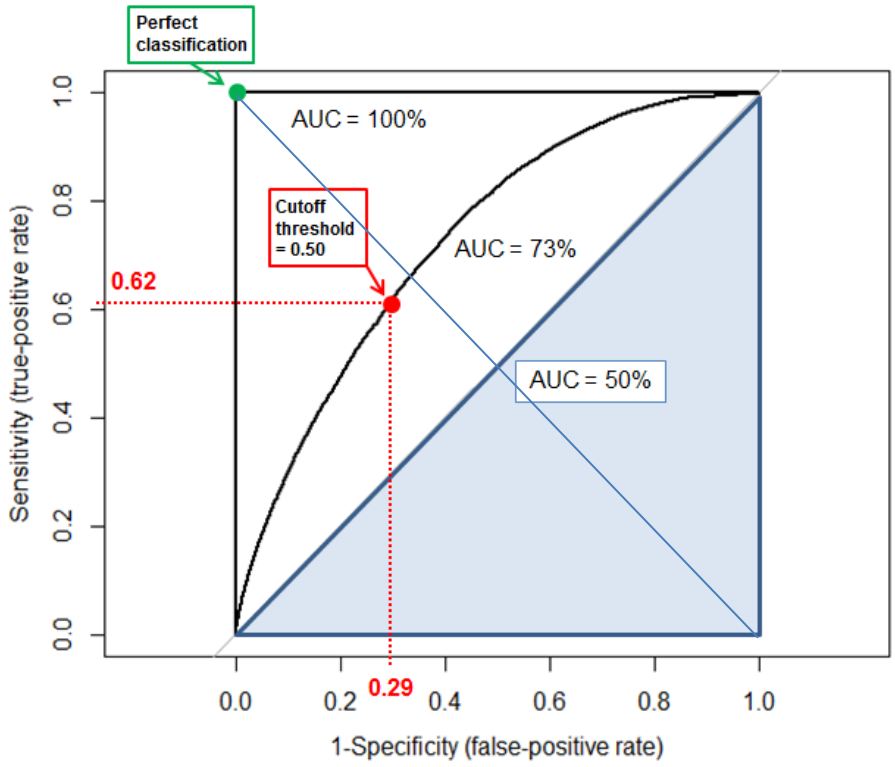

Figure 3: Receiver Operating Characteristic (ROC) curve showing AUC percentage

In the predictive modeling process, different binary classification modeling algorithms will generate different ROC curves. Thus, one could impose ROC curves for each model to visually compare one model versus another, or one could just compare the AUC values that provide a measure of the optimal balance of a high true-positive rate and low false-positive rate across the board [30-32]. We have found that in practice and in the academic literature that the comparison of ROC/AUC is the gold standard binary classification model comparison metric among all possible statistical classification performance measures.

Stochastic Cost-Benefit Analysis

A great motivation for Stochastic Cost-Benefit Analysis is an observation made by Provost and Fawcett [33], *"Even if a model passes strict evaluation tests "in the lab," there may be external considerations that make it impractical. For example, a common flaw with detection solutions (such as fraud detection, span detection, and intrusion monitoring) is that they produce too many false alarms. A model may be extremely accurate (>99%) by laboratory standards, but evaluation in the actual business context may reveal that it still produces too many false alarms to be economically feasible."*

Based on our experience and research of developing models to support the assortment decision, there appears to be some correlation with a model's statistical performance and actual business performance, but there still exists a gap integrating the two intelligently. Moreover, the decision-makers using the decision-support provided by the predictive models are usually not interested in statistical performance but rather financial costs associated to the decisions they make. When a model is built and assessed for a set of products, those products will have certain sales, quality, and feature characteristics that are used to estimate the propensity of it selling. However, one must recognize that the actual direct and indirect costs ($) will vary based on the product. For example, the grocery retailer will often model similar products of the same category collectively. The model might suggest that ketchup brand A has a 55% chance of

selling in a certain location (i.e. “a seller” or “Class 1”), whereas ketchup brand B has a 48% chance of selling in the same location thus a non-seller (i.e. “Class 0”). If brand A is premium and we assume it sells for $3, brand B is the store brand and sells for $1, and both products have the same unit costs, there is an issue of treating these SKUs as equals with regard to statistical assessment. The retailer’s binary classification model might have misclassified these products, but their financial misclassification is not accounted for using the traditional statistical performance measures.

This practical phenomenon naturally leads to motivation for a retailer to assess their predictive model accuracy criteria with non-accuracy-based business criteria which we describe as stochastic cost-benefit analysis. We posit that a retailer building predictive models and using the “best” models to predict the probability that SKU will sell in a particular location is needed. The intuition is for them to be able to identify the most highly desired products to stock in a location, thus leading to the greatest potential sales. However, this logic does not necessarily lead to optimal decision-making and thus achieving the retailer’s true business key performance indicators (KPIs).

Consider the following motivating example that incorporates the sales and costs of each SKU in dollars. A retailer will want to try and minimize the false-positive and false-negatives. Some false-negatives will have a lost sales cost of not stocking the product in a location when it would have sold. False-positives will have their own costs as well, such as inventory, recycling, and lost opportunity costs because the product was purchased and place in a location and it did not sell. In our scenario we only consider the cost of the product and do not incorporate those additional inventory and recycling costs. In the financial form of classification we show here, if the retailer correctly predicted SKU i would sell they would know that the associated profit for that SKU is $p_i = rev_i - cost_i$. However, if SKU i does not sell but was predicted to have sold they would earn $-cost_i$ for making that stocking decision. The total reward of correct and incorrect decisions is shown in **Error! Reference source not found.**.

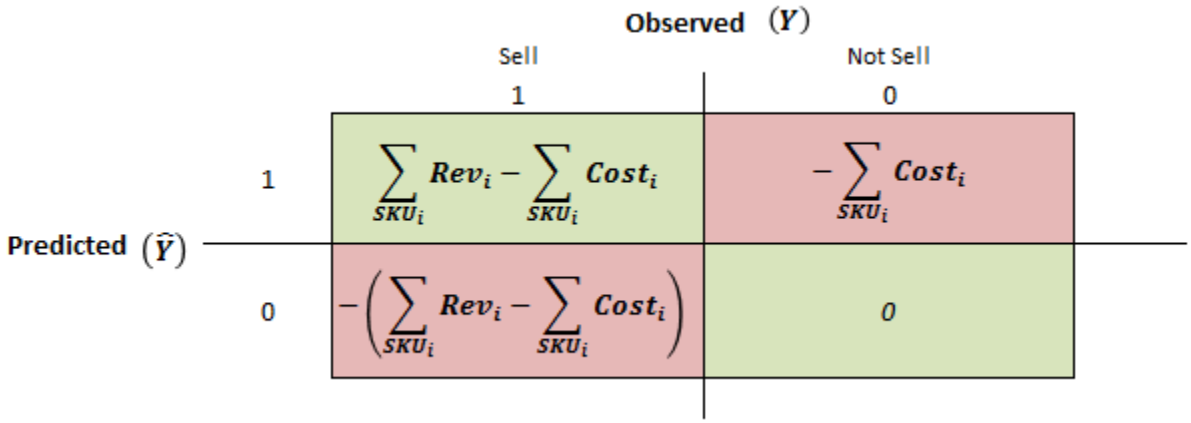

Figure 4: Costs-benefit confusion matrix

If we considered ten products from a similar product category shown in Table 2. These products will have similar domain characteristics, but will have varying prices, costs, margins, and profits. We can see that historically certain products have sold over a specified time frame (i.e. “sellers”) and some have not sold (i.e. “non-sellers”). The model estimates a probability that each product will sell and based on the cutoff threshold of 0.50 each product can be assigned to a class. Using the basic equations in Figure 4, we can calculate the expected return for correctly classifying each SKU.

SKU i	Sales	Cost	Profit	Yhat	Yhat[0-1 | Cutoff=0.50]	Y	Expected Return
A	$15.00	$ 9.00	$ 6.00	0.950	1	0	$ (9.00)
B	$ 1.00	$ 0.46	$ 0.54	0.465	1	1	$ 0.54
C	$ 3.56	$ 1.50	$ 2.06	0.829	1	1	$ 2.06
D	$17.28	$10.00	$ 7.28	0.460	0	1	$ (7.28)
E	$ 5.89	$ 2.00	$ 3.89	0.959	1	0	$ (2.00)
F	$ 9.69	$ 5.00	$ 4.69	0.392	0	0	$ -
G	$ 2.45	$ 1.20	$ 1.25	0.460	0	0	$ -
H	$30.00	$15.00	$15.00	0.620	1	1	$ 15.00
I	$10.22	$ 5.62	$ 4.60	0.320	0	0	$ -
J	$ 3.25	$ 1.49	$ 1.76	0.999	1	1	$ 1.76

Table 2: Example of 10 SKUs expected return based on the decisions made by the predictive model

Figure 5 shows the traditional confusion matrix based on counts of correctly classified and incorrectly classified SKUs. Evaluating the classification accuracy for this model, the overall accuracy is 70% (7/10), sensitivity is 80% (4/5), and specificity is 60% (3/5).

		Observed (Y)		
		Sell	Not Sell	
		1	0	
Predicted $(\hat{Y})$	1	4	2	6
	0	1	3	4
		5	5	10

Figure 5: Confusion matrix based on a decision cutoff criterion of 0.50

Figure 6 shows the cost-benefit of using this model to stock SKUs classified as sellers or non-sellers. In this example, the model would yield a positive net gain, but the misclassification of SKU A has a major overall impact to the business, possibly leading to an overall loss for this product group because the retailer is not incorporating the additional inventory, recycling, and lost opportunity costs.

		Observed (Y)		
		Sell	Not Sell	
		1	0	
Predicted $(\hat{Y})$	1	$ 19.36	$ (11.00)	$ 8.36
	0	$ (7.28)	$ -	$ (7.28)
		$ 12.08	$ (11.00)	$ 1.08
			19.36	$ (18.28)

Figure 6: Estimated business-costs-benefits

If several models were compared in this fashion we posit that more information is gained for decision-making purposes than comparing the models on traditional statistical performance alone. Moreover, model evaluation might lead to better interaction with the decision-makers that the model is supporting, because traditional classification statistics may or may not resonate with them. Based on our experience, talking about expected performance in units they understand (i.e. dollars) has led to much better feedback

and improved decision-support. In theory, had our predictive model displayed better performance, we would have realized a higher business benefit return and lower costs.

RESEARCH DESIGN

To evaluate our proposed stochastic cost-benefit performance measures we tested four different binary classification algorithms (Logistic Regression, Classification tree, C5.0 decision tree, Linear Discriminant Analysis) on one product category from a national retailer. The dataset was randomly split into a 70% training dataset and a 30% out-of-sample testing dataset. Due to the dataset being imbalanced, the training dataset was rebalanced using four commonly used rebalancing techniques (up-Sampling, down-Sampling, SMOTE, and ROSE) as well as evaluated without rebalancing.

We compare the different modeling approaches to identify what the optimal model chosen would have been based solely on traditional binary classification performance measures (e.g. AUC, Overall accuracy) and then compare those results to the stochastic cost-benefit evaluation procedure to see how well they correlate.

For this one product category, this design allows us to answer these three research questions:

1. What are the expected business gains a retailer could expect to achieve when using various combinations of rebalancing and binary classification algorithms to identify SKU sellers from non-sellers?
2. For each rebalance technique/binary classification algorithm combination, can the retailer expect the cost-benefit on the out of sample test set to follow more or less in line with the training data set? Classical statistics do not, as one might have over fit the training dataset.
3. For each rebalance technique/binary classification algorithm combination, can the retailer expect the cost-benefit on the out of sample test datasets to be similar and could using the financial measures be used to identify which model is optimal compared to traditional statistical measures?

RESULTS

As shown in Table 3, each rebalance-model combination yields different traditional classification performance measures, but also cost-benefit business measures. The traditional statistical measures can be compared for all combinations. However, the financial performance can only be compared within its respective rebalance-model group because each set has a different number of observations.

Rebalance	Model	AUC	Accuracy Pct	TP	FP	FN	TN	Expected Profit	Misclassification Costs
Raw	C5.0	0.7553	83.52 (83.18,83.85)	$ 1,159,602	$ (153,131)	$ (68,296)	$ -	$ 938,174	$ (221,428)
	CART	0.7356	81.48 (81.13,81.83)	$ 1,130,527	$ (176,867)	$ (97,371)	$ -	$ 856,289	$ (274,238)
	LDA	0.7892	77.73 (77.35,78.1)	$ 1,202,326	$ (280,017)	$ (25,571)	$ -	$ 896,738	$ (305,588)
	Logit	0.8103	80.49 (80.13,80.84)	$ 1,137,988	$ (182,159)	$ (89,910)	$ -	$ 865,918	$ (272,069)
	Average	**0.7726**	**80.8 (80.45,81.15)**	**$ 1,157,611**	**$ (198,043)**	**$ (70,287)**	**$ -**	**$ 889,280**	**$ (268,331)**
Down	C5.0	0.8427	77.21 (76.64,77.77)	$ 258,798	$ (47,210)	$ (99,993)	$ -	$ 111,596	$ (147,202)
	CART	0.7660	73.55 (72.95,74.13)	$ 247,208	$ (61,222)	$ (111,583)	$ -	$ 74,404	$ (172,805)
	LDA	0.7949	72.05 (71.45,72.65)	$ 198,095	$ (38,931)	$ (160,696)	$ -	$ (1,531)	$ (199,626)
	Logit	0.8080	73.53 (72.93,74.11)	$ 209,329	$ (42,520)	$ (149,462)	$ -	$ 17,347	$ (191,982)
	Average	**0.8029**	**74.08 (73.49,74.67)**	**$ 228,358**	**$ (47,471)**	**$ (130,433)**	**$ -**	**$ 50,454**	**$ (177,904)**
Up	C5.0	0.9133	85.68 (85.43,85.93)	$ 1,020,521	$ (121,283)	$ (207,377)	$ -	$ 691,860	$ (328,660)
	CART	0.7658	73.63 (73.31,73.94)	$ 844,728	$ (212,999)	$ (383,169)	$ -	$ 248,560	$ (596,168)
	LDA	0.7969	72.33 (72.01,72.65)	$ 683,576	$ (135,308)	$ (544,322)	$ -	$ 3,946	$ (679,630)
	Logit	0.8093	73.46 (73.14,73.78)	$ 715,510	$ (147,392)	$ (512,388)	$ -	$ 55,730	$ (659,780)
	Average	**0.8213**	**76.27 (75.97,76.58)**	**$ 816,084**	**$ (154,245)**	**$ (411,814)**	**$ -**	**$ 250,024**	**$ (566,059)**
SMOTE	C5.0	0.8895	81.44 (81.16,81.72)	$ 1,292,446	$ (275,656)	$ (141,012)	$ -	$ 875,779	$ (416,667)
	CART	0.7409	71.57 (71.25,71.89)	$ 1,073,896	$ (352,472)	$ (359,562)	$ -	$ 361,863	$ (712,033)
	LDA	0.7588	69.61 (69.29,69.94)	$ 1,035,752	$ (284,895)	$ (397,705)	$ -	$ 353,152	$ (682,600)
	Logit	0.7893	73.16 (72.84,73.47)	$ 1,031,881	$ (219,117)	$ (401,577)	$ -	$ 411,187	$ (620,694)
	Average	**0.7946**	**73.95 (73.63,74.26)**	**$ 1,108,494**	**$ (283,035)**	**$ (324,964)**	**$ -**	**$ 500,495**	**$ (607,999)**
ROSE	C5.0	0.9621	91.62 (91.37,91.86)	$ 730,055	$ (40,267)	$ (79,993)	$ -	$ 609,795	$ (120,260)
	CART	0.7437	74.17 (73.78,74.56)	$ 418,419	$ (59,269)	$ (391,630)	$ -	$ (32,480)	$ (450,899)
	LDA	0.7603	70.18 (69.77,70.59)	$ 356,813	$ (69,564)	$ (453,235)	$ -	$ (165,986)	$ (522,799)
	Logit	0.7628	71 (70.6,71.41)	$ 399,103	$ (83,621)	$ (410,945)	$ -	$ (95,463)	$ (494,566)
	Average	**0.8072**	**76.74 (76.38,77.11)**	**$ 476,098**	**$ (63,180)**	**$ (333,951)**	**$ -**	**$ 78,967**	**$ (397,131)**
Overall Average		**0.7997**	**76.37 (75.99,76.75)**	**$ 757,329**	**$ (149,195)**	**$ (254,290)**	**$ -**	**$ 353,844**	**$ (403,485)**

Table 3: Training data statistics per rebalance set and model

The C5.0 decision tree model performed consistently the best among all sets, though the Logit model achieved the greatest AUC among all models in the Raw (i.e. not rebalanced) group. What is interesting here is that the expected financial performance measures follow closely with the statistical measures.

When comparing the statistical performance of the training and testing sets in Table 4, we find that all models within the ROSE rebalanced set were over fit.

		Training Data		Testing Data	
Rebalance	Model	AUC	Accuracy Pct	AUC	Accuracy Pct
Raw	C5.0	0.7553	83.52 (83.18,83.85)	0.7561	83.6 (83.09,84.1)
	CART	0.7356	81.48 (81.13,81.83)	0.7383	81.59 (81.05,82.11)
	LDA	0.7892	77.73 (77.35,78.1)	0.7957	77.55 (76.97,78.12)
	Logit	0.8103	80.49 (80.13,80.84)	0.8153	80.68 (80.14,81.22)
	Average	**0.7726**	**80.8 (80.45,81.15)**	**0.7763**	**80.85 (80.31,81.39)**
Down	C5.0	0.8427	77.21 (76.64,77.77)	0.8412	75.27 (74.68,75.86)
	CART	0.7660	73.55 (72.95,74.13)	0.7707	74.55 (73.95,75.14)
	LDA	0.7949	72.05 (71.45,72.65)	0.8014	68.58 (67.94,69.22)
	Logit	0.8080	73.53 (72.93,74.11)	0.8128	70.82 (70.19,71.44)
	Average	**0.8029**	**74.08 (73.49,74.67)**	**0.8065**	**72.31 (71.69,72.91)**
Up	C5.0	0.9133	85.68 (85.43,85.93)	0.9150	85.48 (84.99,85.96)
	CART	0.7658	73.63 (73.31,73.94)	0.7721	74.61 (74.01,75.21)
	LDA	0.7969	72.33 (72.01,72.65)	0.8024	68.61 (67.97,69.25)
	Logit	0.8093	73.46 (73.14,73.78)	0.8136	70.97 (70.34,71.59)
	Average	**0.8213**	**76.27 (75.97,76.58)**	**0.8258**	**74.92 (74.33,75.5)**
SMOTE	C5.0	0.8895	81.44 (81.16,81.72)	0.8546	84 (83.49,84.5)
	CART	0.7409	71.57 (71.25,71.89)	0.7300	77.9 (77.32,78.46)
	LDA	0.7588	69.61 (69.29,69.94)	0.7569	73.16 (72.55,73.76)
	Logit	0.7893	73.16 (72.84,73.47)	0.7866	75.93 (75.34,76.52)
	Average	**0.7946**	**73.95 (73.63,74.26)**	**0.7820**	**77.75 (77.18,78.31)**
ROSE	C5.0	0.9621	91.62 (91.37,91.86)	0.7299	35.06 (34.41,35.71)
	CART	0.7437	74.17 (73.78,74.56)	0.6543	51.15 (50.47,51.84)
	LDA	0.7603	70.18 (69.77,70.59)	0.7995	63.78 (63.12,64.44)
	Logit	0.7628	71 (70.6,71.41)	0.8042	66.77 (66.12,67.41)
	Average	**0.8072**	**76.74 (76.38,77.11)**	**0.7470**	**54.19 (53.53,54.85)**
Overall Average		**0.7997**	**76.37 (75.99,76.75)**	**0.7875**	**72 (71.41,72.59)**

Table 4: Training and testing assessment statistics comparison

This result is interesting and unexpected. Often any particular model might be over fit and will need to be re-specified or re-tuned. In this study, all the models had the same tuning parameters and were trained using the same dataset, but the dataset was rebalanced in a different fashion. Aside from this interesting observation, the C5.0 decision tree performed the best for the down, up, and SMOTE rebalance datasets, and the Logit model performed the best when not rebalancing the training set.

Table 5 shows how the expected profit and misclassification costs for each group based on the out-of-sample testing dataset. Here all rebalance-model combinations can be compared because the dataset consists of the same exact records.

Rebalance	Model	AUC	Accuracy Pct	TP	FP	FN	TN	Expected Profit	Misclassification Costs
Raw	C5.0	0.7561	83.6 (83.09,84.1)	$ 486,671	$ (63,688)	$ (29,452)	$ -	$ 393,531	$ (93,140)
	CART	0.7383	81.59 (81.05,82.11)	$ 470,624	$ (73,996)	$ (45,499)	$ -	$ 351,129	$ (119,495)
	LDA	0.7957	77.55 (76.97,78.12)	$ 504,976	$ (120,780)	$ (11,147)	$ -	$ 373,049	$ (131,927)
	Logit	0.8153	80.68 (80.14,81.22)	$ 477,356	$ (77,276)	$ (38,767)	$ -	$ 361,313	$ (116,043)
	Average	**0.7763**	**80.85 (80.31,81.39)**	**$ 484,907**	**$ (83,935)**	**$ (31,216)**	**$ -**	**$ 369,756**	**$ (115,151)**
Down	C5.0	0.8412	75.27 (74.68,75.86)	$ 369,293	$ (18,835)	$ (146,830)	$ -	$ 203,627	$ (165,665)
	CART	0.7707	74.55 (73.95,75.14)	$ 356,504	$ (25,467)	$ (159,619)	$ -	$ 171,418	$ (185,086)
	LDA	0.8014	68.58 (67.94,69.22)	$ 290,155	$ (16,058)	$ (225,968)	$ -	$ 48,129	$ (242,026)
	Logit	0.8128	70.82 (70.19,71.44)	$ 303,645	$ (17,835)	$ (212,478)	$ -	$ 73,331	$ (230,314)
	Average	**0.8065**	**72.31 (71.69,72.91)**	**$ 329,899**	**$ (19,549)**	**$ (186,224)**	**$ -**	**$ 124,126**	**$ (205,773)**
Up	C5.0	0.9150	85.48 (84.99,85.96)	$ 430,926	$ (14,352)	$ (85,197)	$ -	$ 331,378	$ (99,549)
	CART	0.7721	74.61 (74.01,75.21)	$ 356,921	$ (25,506)	$ (159,203)	$ -	$ 172,212	$ (184,708)
	LDA	0.8024	68.61 (67.97,69.25)	$ 289,662	$ (15,532)	$ (226,461)	$ -	$ 47,670	$ (241,993)
	Logit	0.8136	70.97 (70.34,71.59)	$ 304,229	$ (17,497)	$ (211,894)	$ -	$ 74,838	$ (229,391)
	Average	**0.8258**	**74.92 (74.33,75.5)**	**$ 345,434**	**$ (18,221)**	**$ (170,689)**	**$ -**	**$ 156,524**	**$ (188,910)**
SMOTE	C5.0	0.8546	84 (83.49,84.5)	$ 465,004	$ (46,744)	$ (51,119)	$ -	$ 367,142	$ (97,863)
	CART	0.7300	77.9 (77.32,78.46)	$ 391,443	$ (51,751)	$ (124,680)	$ -	$ 215,012	$ (176,431)
	LDA	0.7569	73.16 (72.55,73.76)	$ 373,992	$ (41,461)	$ (142,131)	$ -	$ 190,400	$ (183,592)
	Logit	0.7866	75.93 (75.34,76.52)	$ 373,971	$ (32,778)	$ (142,152)	$ -	$ 199,041	$ (174,930)
	Average	**0.7820**	**77.75 (77.18,78.31)**	**$ 401,103**	**$ (43,183)**	**$ (115,021)**	**$ -**	**$ 242,899**	**$ (158,204)**
ROSE	C5.0	0.7299	35.06 (34.41,35.71)	$ 47,655	$ (445)	$ (468,468)	$ -	$ (421,258)	$ (468,913)
	CART	0.6543	51.15 (50.47,51.84)	$ 109,901	$ (4,644)	$ (406,223)	$ -	$ (300,966)	$ (410,867)
	LDA	0.7995	63.78 (63.12,64.44)	$ 212,286	$ (10,875)	$ (303,837)	$ -	$ (102,426)	$ (314,712)
	Logit	0.8042	66.77 (66.12,67.41)	$ 243,232	$ (13,491)	$ (272,891)	$ -	$ (43,150)	$ (286,382)
	Average	**0.7470**	**54.19 (53.53,54.85)**	**$ 153,268**	**$ (7,364)**	**$ (362,855)**	**$ -**	**$ (216,950)**	**$ (370,219)**
Overall Average		**0.7875**	**72 (71.41,72.59)**	**$ 342,922**	**$ (34,451)**	**$ (173,201)**	**$ -**	**$ 135,271**	**$ (207,651)**

Table 5: Testing data statistics per rebalance set and model

Our results show that the financial performance measures follow closely in line with the statistical measures. However, without looking at the financial measures the retailer would have likely chosen the Up-C5.0 combination as their final production model to base their decision-support because it has the greatest test statistics (AUC = 0.9150/Accuracy = 85.48%). However, this model does not necessarily lead to the best expected profit or least misclassification costs. In fact, the Raw-C5.0 combination revealed respectable test statistics (AUC = 0.7561/Accuracy = 83.6%) and led to an expected profit gain of $62,153 and reduced misclassification costs of $6,409 compared to using Up-C5.0.

CONCLUSIONS & FUTURE RESEARCH

The results of the study show the value that might be achieved using stochastic cost-benefit analysis for binary classification model assessment and selection. The motivation for this research is that the assortment decision is considered one of the most important decisions that a retailer will make and thus must consider the financial considerations that their stocking decisions could make. Using naïve approaches such as binary classification to identify SKU sellers from non-sellers could be a viable baseline modeling strategy when the number of SKUs to evaluate is in the hundreds of thousands. However, we posit that retailer's should also use more sophisticated approaches that incorporate substitution behavior.

Our study is limited in that only one category of products is investigated in insolation, substitution behavior was disregarded, the evaluation of the business performance is retrospective over the same time frame as the predictions, and modeling build time is not incorporated. Ideally, all these concerns must be addressed by the retailer when generating timely and reliable decision-support.

We want to extend this research by addressing these important financial aspects as well as other realistic concerns that a retailer must face so that complete assortment decision-support system can be created and tested. Factors we are testing include incorporating market basket analysis models, substitution behavior

models, and constructing a valid simulation design. Consumer purchasing baskets frequently contain complementary products from different categories (e.g. ice cream, cake). If one product is left out the assortment it could impact the propensity to purchase and sales of the other item. Substitution-based choice model such as Multinomial Logit are frequently used to estimate propensity to sell among a substitutable set. Understanding the degree of this substitution can provide greater improvements to the retailer's depth, and if ignored can lead to carrying too many similar products that cannibalize each other's sales and lead to additional carrying costs and wasted shelf space. Also, financial considerations could be incorporated as we have shown in the study. The optimal way to estimate truth is to design and run a controlled experiment. While most business studies are observational in nature, business experimentation is being performed more regularly. However, we believe a valid simulation design could constructed by building models using older seasons, employing the decision model that incorporates all these aspects of the assortment decision, and then comparing our results to the latest observed season's sales for products that were stocked. Lastly, to make this study cover all bases, the time to train and score all the various models must be accounted for. We believe that incorporating the previous mentioned items in conjunction with the estimated run time, based on the retailer's resources, would provide a near complete solution.

ACKNOWLEDGEMENTS

The authors would like to thank the retailer, decision-makers, and data scientists we have collaborated with for helping us to better understand the assortment planning process. This project has allowed us an opportunity to identify missing areas in the academic literature, but more importantly improve business practice by following the management science/analytics process.

APPENDIX

Rebalance	Model	AUC	Accuracy Pct	Sensitivity	Specificity	PPV	NPV	Kappa	TP	FP	FN	TN
Raw	C5.0	0.7553	83.52 (83.18,83.85)	0.9618	0.3980	0.8465	0.7513	0.4320	35,892	6,509	1,425	4,304
	CART	0.7356	81.48 (81.13,81.83)	0.9428	0.3733	0.8385	0.6540	0.3729	35,182	6,777	2,135	4,036
	LDA	0.7892	77.73 (77.35,78.1)	0.9755	0.0932	0.7878	0.5245	0.0971	36,403	9,805	914	1,008
	Logit	0.8103	80.49 (80.13,80.84)	0.9455	0.3194	0.8274	0.6296	0.3212	35,285	7,359	2,032	3,454
	Average	**0.7726**	**80.8 (80.45,81.15)**	**0.9564**	**0.2960**	**0.8251**	**0.6398**	**0.3058**	**35,691**	**7,613**	**1,627**	**3,201**
Down	C5.0	0.8427	77.21 (76.64,77.77)	0.7381	0.8061	0.7919	0.7548	0.5442	7,981	2,097	2,832	8,716
	CART	0.7660	73.55 (72.95,74.13)	0.7451	0.7258	0.7310	0.7401	0.4709	8,057	2,965	2,756	7,848
	LDA	0.7949	72.05 (71.45,72.65)	0.6420	0.7990	0.7616	0.6906	0.4410	6,942	2,173	3,871	8,640
	Logit	0.8080	73.53 (72.93,74.11)	0.6840	0.7866	0.7622	0.7134	0.4705	7,396	2,308	3,417	8,505
	Average	**0.8029**	**74.08 (73.49,74.67)**	**0.7023**	**0.7794**	**0.7617**	**0.7247**	**0.4817**	**7,594**	**2,386**	**3,219**	**8,427**
Up	C5.0	0.9133	85.68 (85.43,85.93)	0.8494	0.8642	0.8622	0.8516	0.7137	31,698	5,066	5,619	32,251
	CART	0.7658	73.63 (73.31,73.94)	0.7462	0.7263	0.7317	0.7411	0.4725	27,847	10,213	9,470	27,104
	LDA	0.7969	72.33 (72.01,72.65)	0.6469	0.7996	0.7635	0.6937	0.4466	24,142	7,478	13,175	29,839
	Logit	0.8093	73.46 (73.14,73.78)	0.6841	0.7851	0.7610	0.7131	0.4692	25,529	8,019	11,788	29,298
	Average	**0.8213**	**76.27 (75.97,76.58)**	**0.7317**	**0.7938**	**0.7796**	**0.7499**	**0.5255**	**27,304**	**7,694**	**10,013**	**29,623**
SMOTE	C5.0	0.8895	81.44 (81.16,81.72)	0.9307	0.6594	0.7846	0.8771	0.6090	40,255	11,050	2,997	21,389
	CART	0.7409	71.57 (71.25,71.89)	0.8551	0.5299	0.7080	0.7328	0.3988	36,984	15,250	6,268	17,189
	LDA	0.7588	69.61 (69.29,69.94)	0.7632	0.6067	0.7212	0.6578	0.3736	33,012	12,759	10,240	19,680
	Logit	0.7893	73.16 (72.84,73.47)	0.7850	0.6603	0.7550	0.6973	0.4483	33,954	11,018	9,298	21,421
	Average	**0.7946**	**73.95 (73.63,74.26)**	**0.8335**	**0.6141**	**0.7422**	**0.7412**	**0.4574**	**36,051**	**12,519**	**7,201**	**19,920**
ROSE	C5.0	0.9621	91.62 (91.37,91.86)	0.9102	0.9222	0.9221	0.9103	0.8323	22,038	1,861	2,174	22,057
	CART	0.7437	74.17 (73.78,74.56)	0.6318	0.8529	0.8130	0.6959	0.4841	15,298	3,518	8,914	20,400
	LDA	0.7603	70.18 (69.77,70.59)	0.5817	0.8235	0.7694	0.6604	0.4045	14,083	4,222	10,129	19,696
	Logit	0.7628	71 (70.6,71.41)	0.6253	0.7958	0.7561	0.6772	0.4207	15,139	4,883	9,073	19,035
	Average	**0.8072**	**76.74 (76.38,77.11)**	**0.6872**	**0.8486**	**0.8152**	**0.7359**	**0.5354**	**16,640**	**3,621**	**7,573**	**20,297**
Overall Average		**0.7997**	**76.37 (75.99,76.75)**	**0.7822**	**0.6664**	**0.7847**	**0.7183**	**0.4612**	**24,656**	**6,767**	**5,926**	**16,294**

Figure 7: Model assessment statistics based on the training dataset

Rebalance	Model	AUC	Accuracy Pct	Sensitivity	Specificity	PPV	NPV	Kappa	TP	FP	FN	TN
Raw	C5.0	0.7561	83.6 (83.09,84.1)	0.9593	0.4103	0.8488	0.7449	0.4398	15,342	2,732	651	1,901
	CART	0.7383	81.59 (81.05,82.11)	0.9410	0.3840	0.8406	0.6533	0.3807	15,049	2,854	944	1,779
	LDA	0.7957	77.55 (76.97,78.12)	0.9721	0.0967	0.7879	0.5011	0.0965	15,547	4,185	446	448
	Logit	0.8153	80.68 (80.14,81.22)	0.9461	0.3261	0.8290	0.6367	0.3293	15,131	3,122	862	1,511
	Average	**0.7763**	**80.85 (80.31,81.39)**	**0.9546**	**0.3043**	**0.8266**	**0.6340**	**0.3116**	**15,267**	**3,223**	**726**	**1,410**
Down	C5.0	0.8412	75.27 (74.68,75.86)	0.7378	0.8042	0.9286	0.4705	0.4330	11,800	907	4,193	3,726
	CART	0.7707	74.55 (73.95,75.14)	0.7501	0.7293	0.9054	0.4582	0.3962	11,997	1,254	3,996	3,379
	LDA	0.8014	68.58 (67.94,69.22)	0.6510	0.8060	0.9205	0.4009	0.3363	10,412	899	5,581	3,734
	Logit	0.8128	70.82 (70.19,71.44)	0.6855	0.7865	0.9173	0.4201	0.3604	10,963	989	5,030	3,644
	Average	**0.8065**	**72.31 (71.69,72.91)**	**0.7061**	**0.7815**	**0.9179**	**0.4374**	**0.3815**	**11,293**	**1,012**	**4,700**	**3,621**
Up	C5.0	0.9150	85.48 (84.99,85.96)	0.8534	0.8595	0.9545	0.6295	0.6310	13,649	651	2,344	3,982
	CART	0.7721	74.61 (74.01,75.21)	0.7512	0.7287	0.9053	0.4590	0.3970	12,014	1,257	3,979	3,376
	LDA	0.8024	68.61 (67.97,69.25)	0.6515	0.8055	0.9204	0.4011	0.3365	10,420	901	5,573	3,732
	Logit	0.8136	70.97 (70.34,71.59)	0.6872	0.7872	0.9177	0.4217	0.3627	10,991	986	5,002	3,647
	Average	**0.8258**	**74.92 (74.33,75.5)**	**0.7359**	**0.7952**	**0.9245**	**0.4778**	**0.4318**	**11,769**	**949**	**4,225**	**3,684**
SMOTE	C5.0	0.8546	84 (83.49,84.5)	0.9303	0.5282	0.8719	0.6872	0.4996	14,879	2,186	1,114	2,447
	CART	0.7300	77.9 (77.32,78.46)	0.8567	0.5105	0.8580	0.5079	0.3666	13,702	2,268	2,291	2,365
	LDA	0.7569	73.16 (72.55,73.76)	0.7698	0.5996	0.8691	0.4301	0.3241	12,312	1,855	3,681	2,778
	Logit	0.7866	75.93 (75.34,76.52)	0.7910	0.6501	0.8864	0.4740	0.3897	12,650	1,621	3,343	3,012
	Average	**0.7820**	**77.75 (77.18,78.31)**	**0.8370**	**0.5721**	**0.8713**	**0.5248**	**0.3950**	**13,386**	**1,983**	**2,607**	**2,651**
ROSE	C5.0	0.7299	35.06 (34.41,35.71)	0.1650	0.9912	0.9847	0.2559	0.0773	2,639	41	13,354	4,592
	CART	0.6543	51.15 (50.47,51.84)	0.3950	0.9137	0.9405	0.3044	0.1804	6,318	400	9,675	4,233
	LDA	0.7995	63.78 (63.12,64.44)	0.5785	0.8427	0.9270	0.3667	0.2883	9,252	729	6,741	3,904
	Logit	0.8042	66.77 (66.12,67.41)	0.6256	0.8129	0.9203	0.3861	0.3149	10,005	867	5,988	3,766
	Average	**0.7470**	**54.19 (53.53,54.85)**	**0.4410**	**0.8901**	**0.9431**	**0.3283**	**0.2152**	**7,054**	**509**	**8,940**	**4,124**
Overall Average		**0.7875**	**72 (71.41,72.59)**	**0.7349**	**0.6686**	**0.8967**	**0.4805**	**0.3470**	**11,754**	**1,535**	**4,239**	**3,098**

Figure 8: Model assessment statistics based on the testing dataset

REFERENCES

1. 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.
2. 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.
3. Kök, A.G. and M.L. Fisher, *Demand estimation and assortment optimization under substitution: Methodology and application.* Operations Research, 2007. **55**(6): p. 1001-1021.
4. Sauré, D. and A. Zeevi, *Optimal dynamic assortment planning with demand learning.* Manufacturing & Service Operations Management, 2013. **15**(3): p. 387-404.
5. Fox, E.J., A.L. Montgomery, and L.M. Lodish, *Consumer Shopping and Spending Across Retail Formats*.* The Journal of Business, 2004. **77**(S2): p. S25-S60.
6. Ren, C.R., et al., *Managing product variety and collocation in a competitive environment: An empirical investigation of consumer electronics retailing.* Management Science, 2011. **57**(6): p. 1009-1024.
7. Rooderkerk, R.P., H.J. Van Heerde, and T.H. Bijmolt, *Optimizing Retail Assortments.* Marketing Science, 2013. **32**(5): p. 699-715.
8. staff, C., *Strategic Guide to Big Data Analytics*, C. editors, Editor. 2012, CIO: CIO.
9. Brust, A. *Five Big Data Trends Revolutionizing Retail*. 2013.
10. Andrew McAfee, E.B. *Big Data: The Management Revolution*. 2012.
11. Delen, D. and H. Demirkan, *Data, information and analytics as services.* Decision Support Systems, 2013. **55**(1): p. 359-363.

12. SAS, I. *Big Data Analytics: What it is & why it matters*. 2015.
13. Dwoskin, E., *Big Data's High-Priests of Algorithms: 'Data Scientists' Meld Statistics and Software for Find Lucrative High-Tech Jobs*, in *Wall Street Journal*. 2014: www.wsj.com.
14. Dominic Barton, D.C. *Making Advanced Analytics Work for You*. Harvard Business Review, 2012.
15. Davenport, T.H. *Three big benefits of big data analytics*. 2014.
16. Griswold, M., *Space management: align business challenges and IT vendors.* AMR Research, 2007: p. 1-17.
17. Nelson, E. and S. Ellison, *In a shift, marketers beef up ad spending inside stores.* The Wall Street Journal, 2005: p. A1.
18. Rusmevichientong, P., et al., *Assortment optimization under the multinomial logit model with random choice parameters.* Production and Operations Management, 2014.
19. Gruen, T.W., D.S. Corsten, and S. Bharadwaj, *Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses*. 2002: Grocery Manufacturers of America Washington, DC.
20. Campo, K., E. Gijsbrechts, and P. Nisol, *The impact of retailer stockouts on whether, how much, and what to buy.* International Journal of Research in Marketing, 2003. **20**(3): p. 273-286.
21. Van Woensel, T., et al., *Consumer responses to shelf out-of-stocks of perishable products.* International Journal of Physical Distribution & Logistics Management, 2007. **37**(9): p. 704-718.
22. Ge, X., P.R. Messinger, and J. Li, *Influence of soldout products on consumer choice.* Journal of Retailing, 2009. **85**(3): p. 274-287.
23. Fitzsimons, G.J., *Consumer response to stockouts.* Journal of Consumer Research, 2000. **27**(2): p. 249-266.
24. Hand, D.J. and V. Vinciotti, *Choosing k for two-class nearest neighbour classifiers with unbalanced classes.* Pattern Recognition Letters, 2003. **24**(9): p. 1555-1562.
25. Japkowicz, N. and S. Stephen, *The class imbalance problem: A systematic study.* Intelligent data analysis, 2002. **6**(5): p. 429-449.
26. Menardi, G. and N. Torelli, *Training and assessing classification rules with imbalanced data.* Data Mining and Knowledge Discovery, 2014. **28**(1): p. 92-122.
27. He, H. and E.A. Garcia, *Learning from imbalanced data.* Knowledge and Data Engineering, IEEE Transactions on, 2009. **21**(9): p. 1263-1284.
28. Youden, W.J., *Index for rating diagnostic tests.* Cancer, 1950. **3**(1): p. 32-35.
29. Brennan, R.L. and D.J. Prediger, *Coefficient kappa: Some uses, misuses, and alternatives.* Educational and psychological measurement, 1981. **41**(3): p. 687-699.
30. Altman, D.G. and J.M. Bland, *Diagnostic tests 3: receiver operating characteristic plots.* BMJ: British Medical Journal, 1994. **309**(6948): p. 188.
31. Brown, C.D. and H.T. Davis, *Receiver operating characteristics curves and related decision measures: A tutorial.* Chemometrics and Intelligent Laboratory Systems, 2006. **80**(1): p. 24-38.
32. Fawcett, T., *An introduction to ROC analysis.* Pattern recognition letters, 2006. **27**(8): p. 861-874.
33. Provost, F. and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. 2013: " O'Reilly Media, Inc.".