Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data

BRENT KITCHENS, DAVID DOBOLYI, JINGJING LI, AND AHMED ABBASI

Brent Kitchens (bmk2a@comm.virginia.edu; corresponding author) is an assistant professor of information technology (IT) in the McIntire School of Commerce at the University of Virginia. He holds a Ph.D. in information systems from the Warrington College of Business, University of Florida. His research interests include customer analytics, health IT, and online information dissemination. He has published in various journals, including *Information Systems Research* and *Decision Support Systems*. Before pursuing a career in academia, he worked for five years in IT Risk Advisory at Ernst & Young LLP.

David Dobolyi (dd2es@comm.virginia.edu) is a research scientist in the Center for Business Analytics at the McIntire School of Commerce, University of Virginia. He received his Ph.D. from the University of Virginia, and his primary research interests involve quantitative modeling, experimental cognitive psychology, and artificial intelligence, with applications including cybercrime and health. He has published in numerous journals, including *Science*, and his publications span a broad range of topics among which are reproducibility in science, eyewitness memory, Parkinson's disease, and learning styles theory.

Jingjing Li (jl9rf@comm.virginia.edu) is an assistant professor of information technology in the McIntire School of Commerce at the University of Virginia. She received her Ph.D. from the Leeds School of Business, the University of Colorado at Boulder. Her research interests relate to machine learning and big data analytics, with applications in e-commerce, platform business, health care, search engine, user-generated content, and recommender systems. She received the AWS Research Grant, and Microsoft Research Azure Award for her work on big data analytics. Previously, she was a scientist at Microsoft, where she proposed and implemented large-scale machine learning solutions for Microsoft products such as *Xbox One, Windows 8 Search Charm, Windows Phone App Store, Cortana*, and *Bing Entity Search*.

Ahmed Abbasi (abbasi@comm.virginia.edu) is Murray Research Professor of Information Technology and director of the Center for Business Analytics in the McIntire School of Commerce at the University of Virginia. He earned his Ph.D. from the University of Arizona. His research interests relate to predictive analytics, with applications in online fraud and security, text mining, health, and customer analytics. He has published over 70 peer-reviewed articles in the leading journals and conference proceedings, including *Journal of Management Information Systems, MIS Quarterly, ACM Transactions on Information Systems, IEEE Transactions on*

Journal of Management Information Systems / 2018, Vol. 35, No. 2, pp. 540–574.

ISSN 0742–1222 (print) / ISSN 1557–928X (online)
DOI: https://doi.org/10.1080/07421222.2018.1451957

Knowledge and Data Engineering, and others. His projects on cyber security, health analytics, and social media have been funded by the National Science Foundation. He received the IBM Faculty Award, AWS Research Grant, and Microsoft Research Azure Award for his work on big data. He serves as senior editor or associate editor for several journals. His work has been featured in several media outlets.

Abstract: As more firms adopt big data analytics to better understand their customers and differentiate their offerings from competitors, it becomes increasingly difficult to generate strategic value from isolated and unfocused ad hoc initiatives. To attain sustainable competitive advantage from big data, firms must achieve agility in combining rich data across the organization to deploy analytics that sense and respond to customers in a dynamic environment. A key challenge in achieving this agility lies in the identification, collection, and integration of data across functional silos both within and outside the organization. Because it is infeasible to systematically integrate all available data, managers need guidance in finding which data can provide valuable and actionable insights about customers. Leveraging relationship marketing theory, we develop a framework for identifying and evaluating various sources of big data in order to create a value-justified data infrastructure that enables focused and agile deployment of advanced customer analytics. Such analytics move beyond siloed transactional customer analytics approaches of the past and incorporate a variety of rich, relationship-oriented constructs to provide actionable and valuable insights. We develop a customized kernel-based learning method to take advantage of these rich constructs and instantiate the framework in a novel prototype system that accurately predicts a variety of customer behaviors in a challenging environment, demonstrating the framework's ability to drive significant value.

Key words and phrases: big data, customer acquisition, customer analytics, customer expansion, data integration, data management, design science, IT strategic value, relationship marketing, customer retention.

As companies increasingly compete for customers' attention, raising the cost of acquisition and increasing the difficulty of retention, the strategic challenge of understanding and managing customer relationships has become more difficult and more important [26]. At the same time, it has become easier to obtain vast amounts of data about customers that, when combined with increasingly sophisticated analytical techniques, can provide important and actionable insights [40], allowing companies to innovate and differentiate from competitors. However, with the proliferation of analytics, much of the low-hanging fruit has already been claimed. As such, isolated, ad hoc analytics initiatives can no longer achieve or sustain competitive advantage. Companies that seek to attain strategic value through big data must create focused analytics competencies that provide agility in adapting to a dynamic environment.

With the possible exception of a small number of young, technology-centric companies built directly on analytics capabilities (e.g. Google, Amazon, Capital One), customer analytics have largely been based on siloed data that capture a single aspect of customer behavior at a time [18]. In the past decade, the proliferation and advancement of web traffic analytics, online reviews, social media, customer relationship management (CRM) software, and other information technology-enabled apparatus have caused a rapid expansion in the volume, variety, and velocity of data about customers and their relationships with firms [5, 13]. The firms that will derive the most value from this expansion are those that progress from traditional siloed customer analytics to what we designate as "advanced customer analytics," which integrate a rich variety of relationship-oriented data that enable a deep understanding of customers, driving actionable insights and outcomes for acquisition, retention, expansion, and customer equity.

The integration and use of such rich data sources have great potential for driving competitive differentiation and strategic value; however, there are substantial obstacles to address. Data sources are often unstructured, noisy, and difficult to integrate into a focused and cohesive view of the customer, leaving vanishingly few firms with a coveted "360° view" of customers [34]. Overwhelmed by available and potentially available data, companies must answer many questions, which often fall to information technology (IT) managers [40, 43]: "What data should we collect?" "How do we prioritize data integration efforts?" "How do we quantify the value provided?" and so on. Answers of "just collect and integrate it all" and "we'll figure out the value later" become infeasible as the quantity and variety of available data outpaces the ability to collect, store, and process these data; therefore these decisions must be increasingly informed and intentional [40]. While there have been many calls for guidance in valuing data to inform such decisions [35, 66], there has been little research on the subject. With IT situated as a critical enabler for advanced customer analytics, IT managers are in need of such guidance [40].

The current best practice solution to this problem of data collection and integration is the creation of data lakes: vast repositories where all sorts of data from across an organization are stored in their native format, just waiting to be analyzed and have their value extracted by someone. Data lakes remove barriers and up-front costs for data sharing across an organization, supporting experimentation and discovery. However, the lack of standardization or integration inherent to this approach presents novel challenges [27]. Whenever a new analytics initiative is introduced, data from the data lake must be reorganized from scratch, including the daunting task of fishing relevant data out of the lake and integrating it into a comprehensive view of the customer. More important, if data are not intentionally collected in a way that allows linkages across data sources (e.g., clickstream data with no link to customer IDs), integration may be impossible to achieve with existing data, either stifling the analytics initiative through a lack of relevant data or requiring time-consuming and costly one-off investments in a new data infrastructure. Because this ad hoc approach to data integration and customer analytics requires considerable up-front investment for each new analytics effort, it impairs agility in sensing and responding to customer needs.

In short, a primary challenge in deriving strategic value from big data is the difficulty of creating an integrated big data infrastructure that supports the agile development of advanced customer analytics without overinvestment in worthless data or underinvestment in data that could add significant value. To address this challenge, we follow the design science paradigm to propose a novel framework and system at the intersection of big data analytics, marketing, and IT strategy. Our framework enables the development of advanced customer analytics systems that harness the volume, variety, and velocity of available customer data while assessing the value of various data sources as well as providing much needed guidance for data management and integration decisions and investments.

Based on design principles from relationship marketing theory (RMT), our framework consists of (1) a rich and generalizable set of relationship-oriented constructs that provide insight into customer behaviors; (2) a principled, flexible, versatile predictive model to extract value from a wide variety of structured and unstructured relationship-oriented data; and (3) an approach for estimating the contribution of various constructs for prioritizing data management and integration efforts. Collectively, our framework is focused on building a data infrastructure for agile deployment of customer analytics that leverage and improve customer relationships to drive competitive advantage across a broad range of customer analytics use cases where reliance on siloed data has impeded insight and business value.

We develop a novel kernel-based machine learning method to serve as the predictive model in our framework. This method combines a radial basis function kernel with novel hybrid tree and weighted cross entropy string kernels in a composite kernel support vector machine (SVM). This innovative method allows for the principled embedding of theory and domain knowledge into the constituent kernels; is flexible in incorporating a wide variety of data in tabular, graphical, and text format; and provides versatile ensemble-like performance across a wide portfolio of customer analytics tasks.

We instantiate the framework in a novel advanced customer analytics prototype system developed for a major U.S. e-commerce and catalog-based retailer of educational materials. We evaluate the system and underlying framework using 664,737 actual customers sampled from the firm. The prototype system implements a portfolio of customer analytics applications, accurately predicting customer churn, conversion on specific promotional offers, and lifetime value, all within 30 days of first purchase. Deployment of this portfolio results in significant value for the firm. We also assess the value of various potential data constructs relative to data management and integration costs, offering guidance and justification for investment in data infrastructure that provides agility for deploying further analytics. Because of our work, our corporate partner made significant investments in expansion and integration of data sources found to be most valuable by our prototype, in order to support analytics initiatives.

Our research makes several academic and managerial contributions. Our primary contribution is the synergistic ecosystem of closely related design science artifacts that enable the creation of a value-justified infrastructure for the rapid and agile deployment of a portfolio of advanced customer analytics. Our proposed framework supports development of advanced customer analytics capabilities, directly addressing the challenges of determining which data to invest in and integrate. The novel composite kernel SVM we develop supplies a method tailor-made for extracting insight and value from a rich variety of structured and unstructured relationship-oriented data. The instantiation of our prototype system demonstrates that our framework can be implemented in a working system, providing significant value in a complex real-world environment and demonstrating "last mile" relevance [47]. These artifacts form a valuable contribution to the literature in design science and strategic value of big data.

Our research also has managerial implications for marketers performing customer analytics and IT managers who are asked to support big data analytics initiatives. We provide a general framework for designing advanced customer analytics that keeps both value and costs in mind. As IT managers are asked to prioritize and justify the value of data management and integration activities to support big data analytics [40], the ability to measure the value of each data channel becomes an important consideration [11, 24]. Our broadly applicable framework identifies a variety of novel constructs and provides guidelines for measuring the value of each, supporting data management decisions.

Finally, our research contributes to the literature regarding the use of big data for predictive analytics [5, 13, 22]. Recently, there have been many calls for research regarding the use of big data for decision making to determine whether and how it adds value to organizations. Shmueli and Koppius [56] note a dearth of research examining the value of predictive analytics. Goes [22] specifically extolls the potential benefits of combining a variety of "micro data" from different sources that are enabled by big data analytics. Agarwal and Dhar [5] suggest researchers investigate the resultant ability to study not only organizational and societal but also individual micro-level outcomes. Abbasi et al. [2] point out the need for research that quantifies the value of the volume and variety of data used in big data analytics relative to costs in order to demonstrate effectiveness of big data investments and evaluate the feasibility and efficacy of big data IT artifacts. The current study adds to this nascent literature on big data analytics for prediction, with particular attention to the value of analytics for strategic decision making.

Literature Review

IT Strategy and Support of Big Data Customer Analytics

In order to create sustainable strategic value in a competitive landscape characterized by rapid change, firms must build dynamic capabilities for adapting, integrating, and reconfiguring resources to match their environment [31]. These dynamic capabilities provide strategic agility, which allows firms to quickly recognize and capitalize on opportunities, thereby creating competitive advantage. Sambamurthy et al. [54] demonstrate that IT competencies, when effectively integrated with overall business strategy and capabilities, serve as a platform for agility. However, IT investments to improve agility must be carefully planned and managed, as haphazard or misguided investment can actually impede agility [38].

A burgeoning opportunity for IT to support competitive action exists in the field of big data analytics [12]. In the current environment and for the foreseeable future, analytics represents a primary arena for innovation and competition. The proliferation of abilities to harness big data to improve decisions, processes, and products has made such capabilities a basic requirement for survival, with those best equipped to extract value from data achieving significant competitive advantage [18]. To derive strategic value from analytics, it is important that firms innovate by: (1) moving from general-purpose to specialized analytics uniquely optimized to address specific business issues; and (2) eliminating organizational silos to coordinate data sharing and analytics across functional boundaries [50].

Achieving these two objectives is prohibitively expensive and time-consuming through ad hoc efforts. Instead, IT must strategically partner with other business functions and become a proactive advocate and architect for analytics [57]. Specifically, IT departments should provide data governance and infrastructure to support agile integration of data from multiple sources, organizing "around data as if it were a valuable organizational asset" [50, p. 15] to foster innovation and sustained competitive advantage from big data analytics [66]. By creating an infrastructure that incorporates data across functional silos, "resulting client services are superior, *less susceptible to commoditization*, and generate higher revenue" [39, p. 217 (emphasis added)]. If achieved, this structure can provide a foundation for establishing a portfolio of specialized yet coordinated analytics initiatives that deliver strategic value.

A key challenge relates to the collection and integration of valuable data across silos within and outside the organization. Data management has long been considered a cornerstone of the IT function [23], and "big data's rise has further amplified the importance of IT in this role" through challenges and opportunities of exponentially increasing data volume, variety, and velocity [2, p. 2]. Together these aspects bring into focus the need for data infrastructure investment as well as the potential value of resulting analytics [5, 13, 22]. However, unfocused data management and integration can be extremely costly, and benefits are not always sufficient to offset these costs [24]. Research suggests that "no single integration strategy is optimal in all cases" [11, p. 89], and an intermediate level of integration is often more beneficial than complete integration [11, 24, 40]. In order to support big data analytics through an integrated data infrastructure, organizations must find effective strategies to assess the value of available data.

Given the wide variety of available sources of relevant data, customer analytics initiatives stand to gain significantly from such value-driven investments in IT infrastructure for integration. By building on one of the firm's most important resources (its customers) and one of its least imitable (its data), customer analytics represent an important strategic initiative with the potential to create significant and sustainable competitive advantage. Customer analytics are enabled by a firm's customer agility—the ability to sense opportunities for innovation and respond to those opportunities with competitive action—and operational agility—the ability to rapidly redesign processes to exploit marketplace conditions [54]. These dynamic capabilities should be supported by the synergistic combination of interfunctional business coordination and IT infrastructure for integrating the right data across the organization [53].

As firms move toward analytics specialization, there are opportunities to create a diverse portfolio of customer analytics initiatives spanning acquisition, retention, and expansion in order to optimize customer lifetime value and equity [26]. In order to be most effective, this portfolio of specialized initiatives should draw from various aspects of the organization to incorporate the data most valuable for accomplishing each individual objective. For this to become feasible, a common framework is needed for designing customer analytics applications that incorporates business and IT strategy. With this, firms could build a value-justified infrastructure for supporting a portfolio of advanced customer analytics capabilities that combine to create significant strategic value and sustainable competitive advantage.

Customer Acquisition, Retention, and Expansion

Analyses across the customer lifecycle—encompassing acquisition, retention, and expansion—have become a critical focus for firms as many shift from a historically product-centric orientation to one that is more customer-centric, with emphasis on retaining and building profitable relationships with customers [26]. Current approaches to examining customer retention may largely be traced back to Schmittlein et al. [55], who focus on simple patterns in customer purchase activity to estimate the probability that a customer is still “alive” (i.e., will continue to repurchase in the future). Using only the observed recency, frequency, and monetary value (RFM) of purchases, the model provides probabilistic predictions for individual customer churn. This work originated a stream of research that has had much success at generating parsimonious models that are practical for accurately predicting behaviors for individual customers [20]. However, customers must have substantial purchase histories for the models to be effective, and thus the settings considered by this research are limited to businesses with relatively high purchase volumes. Outside of these high-volume environments, the models lose predictive power.

Aside from the RFM paradigm that dominates customer analytics in marketing literature, understanding customer behavior has become a common use case for the application of machine learning, with many different techniques employed. Neslin et al. [46] were among the first to describe machine learning for customer analytics in a paper summarizing the results of 44 entries to a contest for predicting customer retention, determining which methodological choices impact the quality of prediction. Subsequent studies have applied various methods of all degrees of sophistication [16, 36, 46]. However, these studies focus on the machine learning methods themselves, largely agnostic of the specific setting of customer analytics. Most completely omit details of the setting and data set, with no description of how customer behaviors were measured, what features were included, or how the results could be used in context. These studies contribute to knowledge regarding model selection aspects of customer analytics without regard to other inputs to the analytics process. Of the studies that do provide details, many rely heavily on transactional RFM constructs. For instance, Ballings and Van den Poel [6] employ demographic and transactional features to predict churn. They show that reasonable model performance with these features can require five or more years of data for training. But given the importance of timeliness in the pursuit of strategic business value [13, 40], there is a strong need for relationship-oriented analytics that can produce results in weeks rather than months or years.

Research Gaps

From our analysis of relevant prior literature, we identify three important research gaps. First, there is a lack of research providing guidance in evaluating the impact of data management capabilities at a granular level that could inform prioritization decisions for focused data management and partial integration, which has been called for in the literature [11, 24]. While some studies have focused on the strategic value of data management capabilities at the organizational level or benefits versus related costs of collecting and integrating certain data for specific purposes [43], none provide a general framework for valuing data for analytics initiatives in a broader sense. While there have been calls for this type of research, little has been undertaken. This ability to value various data sources is imperative for developing an integrated big data infrastructure for the agile development of analytics.

Second, there is a lack of studies that have taken a comprehensive and holistic relationship view, as opposed to a siloed transactional view, in predicting customer behaviors. As discussed, prior studies have largely focused on RFM [20, 55] or practically *whatever data are available* [36, 46]. There has even been criticism of this approach, heretofore unaddressed, pointing out that a vast majority of customers at many firms have made only a single purchase and are consequently largely indistinguishable from an RFM perspective, rendering the approach useless [41]. As we will discuss later, many studies have evaluated relationship-oriented constructs associated with customer behaviors using explanatory models [7, 17], yet few have been utilized for prediction, and few employ a multifaceted perspective, whether in a predictive or explanatory context. A more holistic and integrated view of customer relationships would provide more accurate and broadly applicable analytics.

Third, there is a paucity of literature on how to formally operationalize the relationship-oriented constructs and data valuation appraisal alluded to in the first two gaps for the purposes of performing predictive customer analytics. From a design science perspective, instantiations offer essential prescriptive guidelines and proofs-of-concept regarding how IT artifacts might be developed to solve an important class of problems [28]. Such instantiations move beyond conceptual claims and offer practical, research-grounded operationalizations that not only advance the literature but also inform practice.

In this study, we address these gaps. We present a framework specifically focused on providing a structure for determining the value of individual data sources, both alone and in combination, in order to help IT managers decide where to invest resources and how to justify data management and integration efforts in support of advanced customer analytics. The framework includes the identification of a rich set of relationship-oriented constructs for use in advanced customer analytics systems. Our prototype system and novel machine learning method, evaluated in the live setting of an e-commerce retailer, demonstrates the significant value that analytics developed using this framework and construct set may provide.

Framework for Advanced Customer Analytics

The design science research paradigm provides principles for the creation of innovative IT artifacts that allow organizations to improve capabilities, address ever more challenging issues, and take advantage of new opportunities, including those afforded by big data [28, 64]. We propose a framework for implementing advanced customer analytics solutions, including a rich set of constructs that may be used to describe and predict consumer behavior. We also create a novel kernel-based learning method to extract value from this rich variety of structured and unstructured data. Using the framework and method, we instantiate a prototype customer analytics system in a challenging environment requiring rich, relationship-oriented data to make accurate predictions not possible using siloed, transactional data.

Under the design science paradigm, kernel theories are often used to inform the development of novel design products [64]. In the design of our framework, method, and instantiation, we primarily reference the theory of relationship marketing. Relationship marketing theory (RMT) frames interactions with customers not as a set of discrete transactions, but as a relational exchange. Relationships with customers provide a source of sustainable competitive advantage and strategic value for the firm [48]. Combining customer perceptions, preferences, interactions, and communication via various channels, this humanized view of customers stands in contrast to the mechanical view of traditional customer analytics.

RMT has specifically informed a body of literature focused on understanding how perceptual constructs such as satisfaction and repurchase intentions influence customer behaviors through explanatory rather than predictive models. Crosby and Stephens [17] provide some of the earliest research representative of this stream. Surveying a sample of customers, they measure otherwise unobservable opinions and perceptions, including various aspects and antecedents of satisfaction with the purchase (in this case, whole life insurance). They show that more satisfied customers are more likely to renew or upgrade their policies with the same company. Much research has followed this approach, using surveys to measure perceptual constructs related to many aspects of customer relationships with firms and assess how they influence future behavior. Beyond satisfaction, studies have examined the impact of service failures, demographics, service usage, and attachment styles on customer retention [7, 14, 44].

RMT provides design principles that guided all aspects of our proposed framework (see online Appendix A for detail). The relationship marketing literature is extensive and varied but has a distinct focus on understanding customer relationships as opposed to predicting them. In addition to implementing modeling techniques geared toward explanation rather than prediction, many studies include perceptual constructs operationalized through

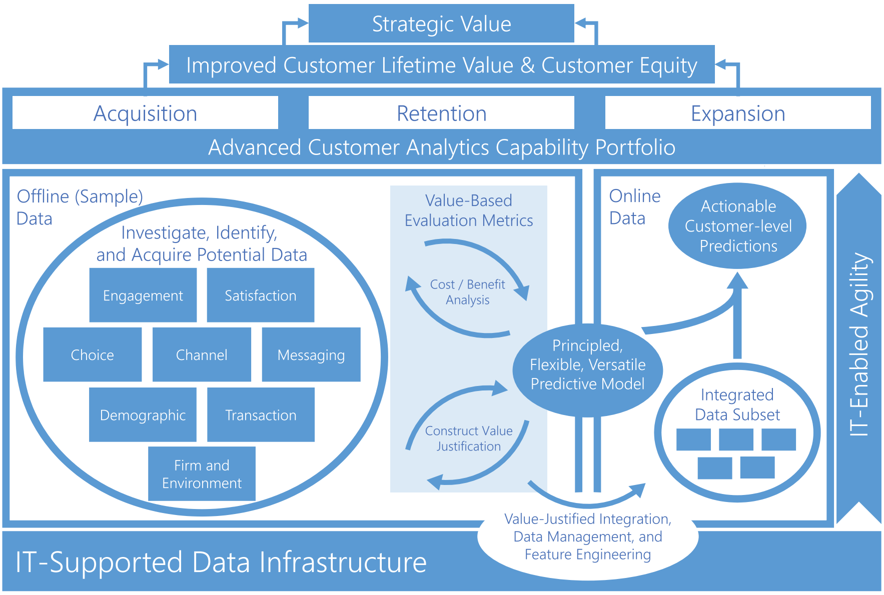

Strategic Value
Improved Customer Lifetime Value & Customer Equity
Acquisition
Retention
Expansion
Advanced Customer Analytics Capability Portfolio
Offline (Sample) Data
Investigate, Identify, and Acquire Potential Data
Engagement
Satisfaction
Choice
Channel
Messaging
Demographic
Transaction
Firm and Environment
Value-Based Evaluation Metrics
Cost / Benefit Analysis
Construct Value Justification
Principled, Flexible, Versatile Predictive Model
Online Data
Actionable Customer-level Predictions
Integrated Data Subset
IT-Enabled Agility
Value-Justified Integration, Data Management, and Feature Engineering
IT-Supported Data Infrastructure

Figure 1. Advanced Customer Analytics Framework

surveys, which makes implementing the resultant models for individual-level predictions difficult. However, the concepts introduced in this literature are informative in building predictive analytics for nuanced problems requiring a holistic and integrated view of customer relationships from a variety of sources. Figure 1 depicts the proposed framework, including:

1. A rich set of relationship-oriented constructs to guide the identification and acquisition of valuable data for advanced customer analytics
2. A principled, flexible, versatile predictive model for extracting value from data constructs
3. Value-based evaluation metrics for action- and outcome-oriented cost/benefit analysis
4. Construct evaluation leading to value-justified integration and data management investments in IT infrastructure to create a foundation of integrated data for analytics
5. A portfolio of advanced customer analytics capabilities supported by IT-enabled agility for providing strategic value through enhanced customer lifetime value and customer equity

Advanced Customer Analytics Constructs

In order to guide the identification and acquisition of valuable data sources, we develop a comprehensive and generalizable set of potential construct categories beneficial for advanced customer analytics. Many of these have rarely been used to predict customer behavior due to the difficulty of integration. The synergistic value of integrated data from a rich variety of sources is the key benefit of advanced customer analytics. To identify the constructs for consideration in advanced customer analytics systems, we performed a broad search of the relationship marketing literature.[1] Here we briefly describe each of the identified constructs and the rationales for their inclusion.

Transaction: Transaction characteristics represent the most basic construct in regard to customer relationships and form a basis for any customer analytics solution. As discussed, the most popular approaches for predicting customer behaviors rely solely on measurements of the recency, frequency, and monetary value (RFM) of customer transactions [20,55]. Transaction characteristics may include useful details in addition to RFM, such as discount levels, promotions, payment types, and specific items purchased, all of which may provide valuable information about customer characteristics and behaviors.

Demographics (and Intangible Customer Characteristics): Demographics are also among the most basic of characteristics firms can measure about their customers. Relationship marketing studies have specifically found that demographic characteristics influence customers' baseline expectations and thresholds for other constructs such as satisfaction, influencing repurchase intentions and behavior [44]. Other studies have specifically focused on how demographics impact customer behavior or simply incorporate demographics as basic features in models focused on other data sources [36, 52, 59]. The combination of transaction and demographic characteristics form what we consider the basic starting point for designing a customer analytics solution. Other intangible customer characteristics may influence relationships with a firm, such as expertise, perceived risk/risk affinity, and technical self-efficacy [8].

Engagement: From an RMT perspective, customers may engage with firms in many ways aside from purchase transactions, with the quantity and types of engagements offering important clues regarding future behavior [19]. A customer may visit a website, open an e-mail, call customer service, post a review or opinion, participate in a forum, and so forth. Engagement provides a signal of a customer's interest in a continuing relationship with the firm. Relationship marketing research has often focused on customer service interactions related to service failures and recoveries [7, 25]. Other studies include basic constructs related to engagement, such as usage patterns, loyalty program and "wish list" usage, and e-mail opt-in/opt-out [62]. There is much variability in the sophistication with which firms are able to collect, measure, and analyze engagement interactions, particularly through outlets not controlled by the firm (e.g., social media), which may have led to the dearth of research investigating its value [19, 63]. As these capabilities improve, customer engagement has potential as a rich resource for advanced customer analytics. A key challenge with engagement data is integrating it with other customer data for analysis. Our framework provides a structure for identifying and valuing such information, supporting integration efforts.

Satisfaction (and Related Perceptions): As discussed previously, customer satisfaction is a key, foundational construct in RMT [7]. Many studies have shown that increasing satisfaction leads to better customer retention and higher customer lifetime value [7, 44]. Beyond satisfaction, other intrinsically related perceptual constructs studied include trust, commitment, loyalty, and other perceptions of firm and environmental attributes [61]. The key challenge with satisfaction and related perceptions is that they can be difficult to measure for individual customers. Most studies examine these perceptions on a sample basis through surveys, with the intent of explaining customer relationships or measuring entire market share rather than making individual-level predictions [17, 44].

Choice: A key observation available from a customer's first as well as subsequent purchases is the choice of product(s) purchased. Particularly when a firm sells goods or services that are horizontally differentiated over differing customer preferences, understanding the types of products a customer prefers can be useful in several ways. From an RMT perspective, some have argued that the categories and types of products purchased by a customer can be an indicator of their level of trust and interest in a particular company [42]. The amount of variety in products purchased may also provide a signal, with cross-buying increasing switching costs as consumers become more aware of the firms offerings and quality [52]. Product choice can also be informative regarding what products the consumer would be interested in purchasing in the future, and product assortment is an important factor in retaining customers.

Channel: The channel(s) through which customers are acquired and continue to interact with a firm provides information about customers and sets the stage for ongoing relationships. RMT suggests that various channel interactions may impact the loyalty or connection customers feel to the firm [8]. For instance, a customer calling and speaking to a representative may develop a stronger relationship with the firm as a result of this communication. Studies have also suggested that customers acquired through digital channels tend to be more loyal and active due to self-selection effects and greater opportunities to form connections with the company [29]. In addition, each channel is likely to attract a different type of customer [32]: for example, online channels may attract those who are more technology-savvy and hedonic.

Messaging: From an RMT perspective, the communications the customer receives from the firm can also have a profound impact on future behavior [17]. Messaging can often be focused specifically on selling: for instance, by using promotional offers to entice a purchase. However, relationship-building messaging specifically focused on enhancing the customer's view of the firm is effective at retention, yet promotional messaging, while effective in the short term, can have various effects over time [21]. The mode and quantity of communications received may also impact customer behavior, with both too little and too much communication being detrimental [59].

Firm and Environment Characteristics: Characteristics of the firm and the market environment in which it operates set the stage for the customer's relationship with the firm. Brand equity, payment equity/perceived fairness of the firm's pricing policy, and firm ethics and citizenship have all been shown to play a significant role in customer perceptions and relationships with the firm [61]. Outside the firm, customer relationships may be influenced by characteristics of the environment, such as dynamism, munificence, complexity, competition and characteristics of competing firms, and market share [8, 62]. These represent important features for incorporation into analytics applications, but also importantly should inform the entire analytics process. Particularly, changes in characteristics or perceptions of the firm and environment should be monitored in order to update analytics solutions for continued relevance.

Principled, Flexible, Versatile Predictive Model

Relationship-oriented data can be nuanced and take on many different formats, both structured and unstructured. Yet, traditional machine learning methods are limited in the complexity of data they can analyze. In order to ensure full extraction of available value within each construct category for producing customer insights, the framework should employ a predictive model that is:

- Capable of principled/theory-driven prediction to exploit the nuances of relationship-oriented data
- Flexible in accommodating complexity of a wide variety of data constructs of various formats
- Versatile in application to a variety of customer analytics initiatives

Kernel-based learning methods embody each of these characteristics, allowing theory embedding via customized kernel functions, incorporation of structured and unstructured data, and adaptability to a wide range of tasks [1, 15], making it possible to effectively represent the complexities of customer behavior.

Value-Based Evaluation Metrics

Aside from identification of a comprehensive set of constructs—the variety of which provide a basis for performing advanced customer analytics—a key contribution of our proposed framework is the presentation of guidelines for directly measuring the value of an analytics solution, both in whole and relative to its constituent parts. This is a necessity for advanced customer analytics endeavors, as essential support for management of the volume, variety, and velocity of required data can be costly and must be justified. Synthesizing the nascent literature on the value of big data analytics [40, 51] and infonomics [35], the key steps for creation of value-based evaluation metrics follow.

Identify Potential Action(s): An intermediate outcome of the final advanced customer analytics product will be predictions regarding customer behavior. But in isolation predictions are useless: only when predictions inform decisions to take action do they have the opportunity to create value for the organization [51]. These decisions and actions enabled by the system are what ultimately provide actual value to the organization through better relationships with customers, influence on customer purchase behaviors, reduced costs, and so on. It is therefore critical that any advanced customer analytics solution has these final value-driving decisions and actions as an ultimate goal. Thus, the first step in measuring the value of the advanced customer analytics solution is to determine potential actions to be taken once predictions are made. In the case of churn prediction, one potential action might be to eliminate all marketing efforts with positive marginal costs to customers predicted to have a high churn propensity [52]. Alternatively, a targeted retention campaign might be created to attempt to convert those with high churn propensity into returning customers [16, 36]. Various possibilities exist. The key is to strategically consider options before and during the design of the advanced customer analytics application, incorporating the potential actions into evaluation of the solution.

Determine Action Value: Once the potential actions to be taken on predictions are known, a value must be determined for taking the proposed action on each customer. This process is similar to constructing a cost matrix for cost-sensitive classification models [1]. A cost or benefit must be assigned to correct and incorrect positive and negative predictions. In the churn case, assuming the potential action is cessation of marketing activities to likely churners, costs and benefits to be considered include: (1) the marginal cost of marketing to a customer, which could be saved for those customers predicted to churn; and (2) the expected revenue to be gained from a customer who continues to make purchases, which may be lost for those who stop receiving marketing materials [52]. Determining the values for potential actions will require assumptions based in part on analysis of historical activities [1].

Define Cutoffs: With potential actions and related values determined, the final step before implementing the value-based evaluation metric for measuring competing models is to define which customers will receive action. This need not be the entire set of customers, and in many cases, particularly with imbalanced classes, may represent a small proportion of the whole [56]. In the example given previously, marketing was to be discontinued to those customers with a high churn propensity. Any model implemented in an advanced customer analytics solution will only provide an estimate of the likelihood of a certain outcome for each customer. Careful consideration must be given to how to determine what share of customers will receive specific potential actions [52, 60].

Generate Final Metric: The actions, value, and cutoffs determined in this step are combined into a complete value-based evaluation metric, which is critical in later steps to evaluate constructs and models, guide data management, integration, and real-time feature engineering efforts, and demonstrate potential return on investment for marketing and IT managers to justify their efforts. For each combination of potential action and actual outcome, a net cost/benefit is defined and applied to all predictions. In order to arrive at a total expected value, the per customer value is multiplied by the number of customers in the live population who exceed the cutoff. The value for an entire analytics portfolio may be calculated by aggregating the total expected value for each individual application. This sum may be used for evaluating and justifying investments in infrastructure to support advanced customer analytics capabilities.

Construct Evaluation and Value-Justified Data Management, Integration, and Feature Engineering

As mentioned, the cost of data management, integration, and real-time feature engineering is among the most important of considerations in any big data analytics initiative [11, 40]. A key aspect of the proposed advanced customer analytics framework relates to evaluating each potential construct and data source with regard to the value-based metrics identified earlier, allowing IT departments supporting advanced customer analytics initiatives to focus efforts on integrating data from only the most valuable sources. In order to quantify the individual value of each construct, we propose the use of an add-in/leave-out approach. Comparisons may be made for each individual construct by running models excluding it from the full model with all constructs and also by adding it to the base model with only the most basic constructs available for production use without additional investments in data management and integration. Additional models may be run with subsets of constructs deemed to contribute significant value. The result of this analysis provides input for final decisions regarding which constructs and data sources should receive investments in data management, integration, and real-time feature construction. The value of each construct and data source may be compared to the effort required in these areas to provide inputs for a final production model and may be used to justify related expenditures and identify the most valuable sources for partial integration [11, 24]. This results in value-justified investments in IT infrastructure supporting agile deployment of advanced customer analytics.

Portfolio of Advanced Customer Analytics

The ultimate goal of the framework is to enable the design of a portfolio of advanced customer analytics applications supported by this value-justified IT infrastructure and the agility it enables. This portfolio should include applications ranging across customer acquisition, retention, and expansion. By enhancing customer relationships across these dimensions, the portfolio can be used to improve customer lifetime value and customer equity, creating strategic value and sustainable competitive advantage [26].

Prototype System Instantiation

Based on the proposed framework for advanced customer analytics, we develop a prototype system to show that the framework can feasibly be implemented in a working system [28]. In order to demonstrate suitability to our framework's intended purpose of supporting the creation of advanced customer analytics, we must identify an application representative of this circumstance. The defining feature of advanced customer analytics is the ability to provide predictive insights into problems not solvable using a traditional siloed view, but instead requiring the support of rich, relationship-oriented data integrated from a variety of sources. Despite the fact that a significant amount of research has been produced regarding customer behavior prediction, the transaction-oriented methods provided require long customer histories (often five or more years) and are less accurate and effective outside high-purchase-volume environments [6]. Many firms provide goods or services that are purchased infrequently by nature or attract a large number of single-purchase customers, and no existing method addresses this challenging case. We propose that advanced customer analytics based on rich, relationship-oriented customer data will provide an effective solution in this setting where other methods fall short.

To create our prototype system, we partnered with a company that we will refer to as Course Shop International (CSI), a large e-commerce and catalog-based seller of educational materials for lifelong learners. CSI is highly representative of the high single-purchase, low-frequency environment of interest. A large majority of CSI's customers make only a single initial purchase, and of those who return, many months may pass between purchases. The goal of our prototype system is to make predictions about customers' future behaviors just 30 days after their initial purchase, when these predictions are most valuable. During this period, we observe aspects related to each construct in our framework. An illustration of the problem setting for our prototype system is provided in Figure 2. To demonstrate how firms may use our framework to develop a data infrastructure providing agility in supporting a portfolio of analytics initiatives, our prototype system consists of three distinct customer analytics applications:

- A churn prediction application focused on evaluating which customers the firm should invest in through continued marketing efforts (retention)
- A conversion prediction application for identifying customers likely to respond to individual email promotions to reduce messaging fatigue and prevent attrition (retention/expansion)

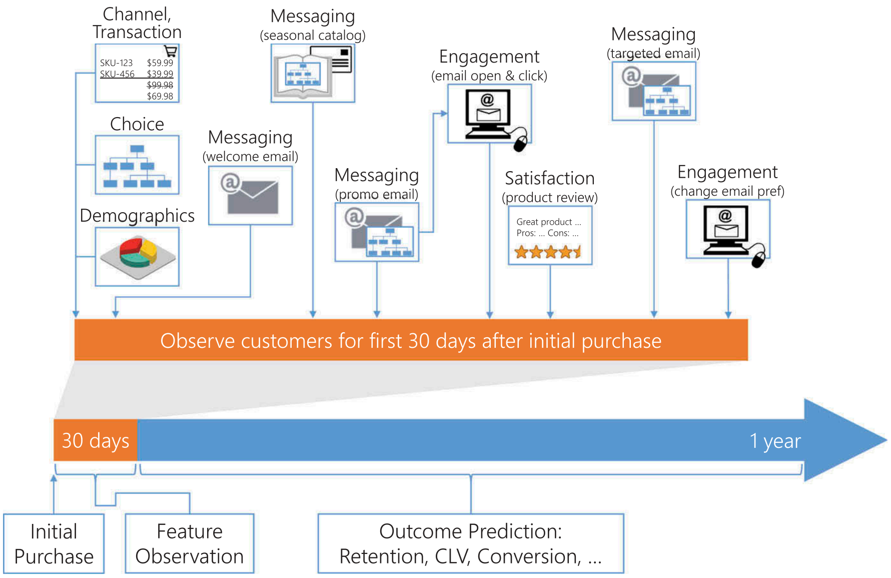

Channel, Transaction
SKU-123 $59.99
SKU-456 $39.99
$99.98
$69.98
Choice
Demographics
Messaging (welcome email)
Messaging (seasonal catalog)
Messaging (promo email)
Engagement (email open & click)
Satisfaction (product review)
Great product ... Pros: ... Cons: ...
Messaging (targeted email)
Engagement (change email pref)
Observe customers for first 30 days after initial purchase
30 days
1 year
Initial Purchase
Feature Observation
Outcome Prediction: Retention, CLV, Conversion, ...

Figure 2. Problem Setting: Customer Agility through Advanced Customer Analytics

- A customer lifetime value (CLV) prediction application for identifying customers who could successfully be expanded through offer of participation in a premium loyalty program (expansion)

These three applications allow us to evaluate each aspect of our framework and the strategic value it creates. Here we focus on the churn prediction example. Details of Results of the combined portfolio are discussed in the evaluation section, while details of the CLV and conversion prediction tasks are provided in online Appendix B.

Leveraging the proposed advanced customer analytics framework described previously, we implemented a novel churn prediction system for CSI. Figure 3 shows the system diagram, encompassing five stages: *Data Lake, Feature Generation, Data Preparation, Offline Modeling/Evaluation*, and *Online Implementation*. These five stages are closely aligned with facets of the proposed framework. For instance, the *Data Lake, Feature Generation*, and *Data Preparation* components of the system relate to the *Investigate, Identify, and Acquire Potential Data* section of the framework. *Offline Modeling/Evaluation* is associated with *Predictive Modeling* and *Value-Based Evaluation Metrics*. Lastly, the *Online Implementation* component relates to the framework's deployment-oriented *Online Data* portion.

Data

Data Lake: For use in constructing the prototype system, we obtained a variety of data for a sample of customers making initial purchases between January 1, 2012, and March 1, 2014. As previously alluded to, incorporating rich relationship-oriented constructs requires consideration of an array of structured and unstructured data sources: a nontrivial task [13]. The various data sources incorporated in the system include structured data from databases that support online transaction processing (OLTP) and CRM, as well as unstructured data in the form of text log files, call center transcripts, and so on, which are collectively referred to as a "data lake." The *Data Lake* was generated by obtaining these various raw data from CSI for a sample of 664,737 customers. As depicted in Figure 4, the data lake included over 188 million raw data points.

Feature Generation: Once the data lake was constructed, the *Feature Generation* component of the system was used to operationalize relationship-oriented constructs identified in the framework. In order to simulate a realistic environment for implementation of the prototype system and avoid data leakage that could inflate accuracy, we utilize a chronological rolling-window approach for creating training and test sets. Only the first 30 days of customer history after initial purchase was used to construct input features, and a period of 365 days to observe outcomes. In total, we generated 1,003 features pertaining to the various construct categories. *Transaction* and *Demographics* represent readily available baseline constructs. *Transaction* features included order timestamps, amounts, prices, discounts, payment and shipping methods, as well as purchased product information such as course names and topics.

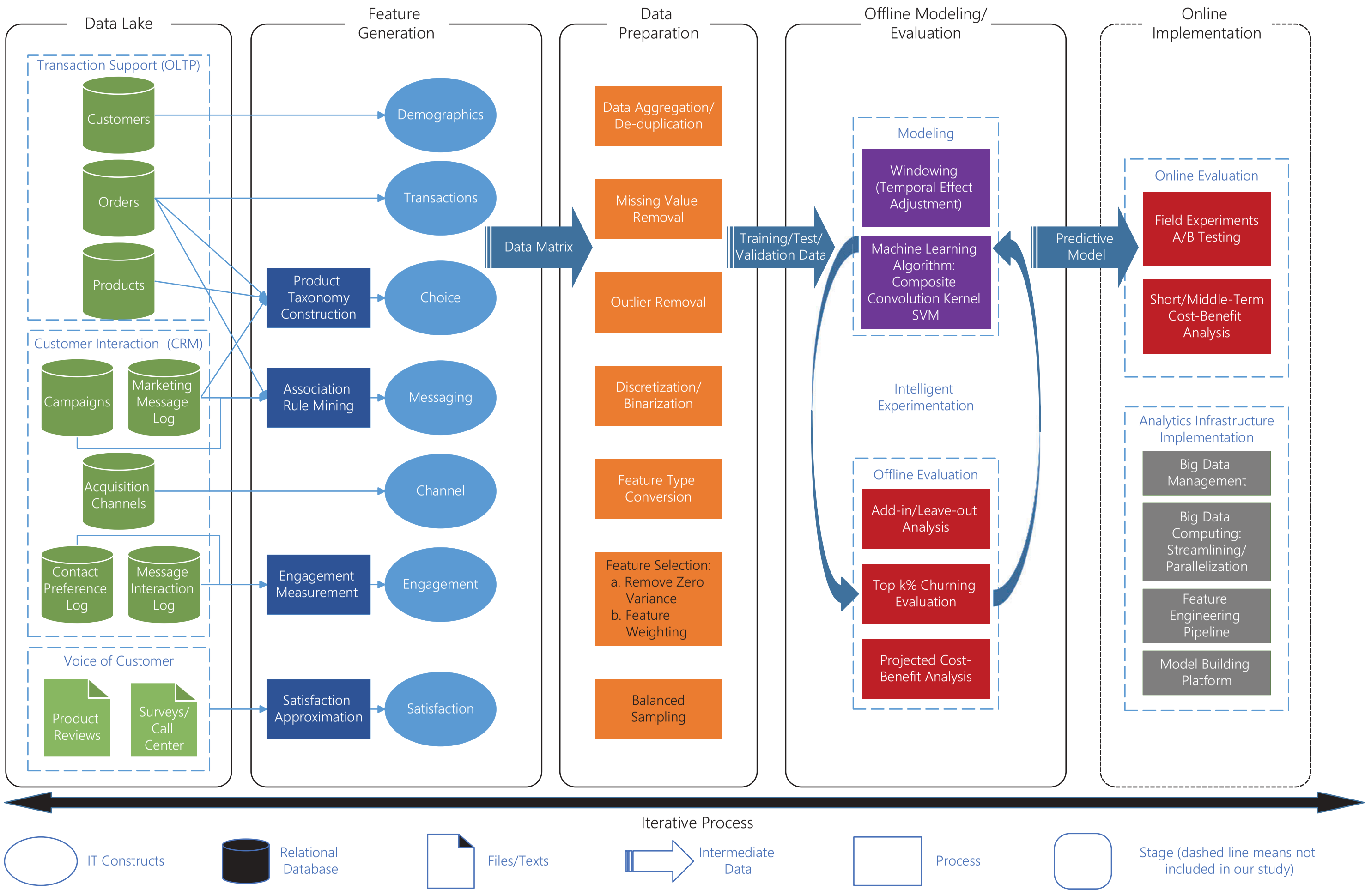
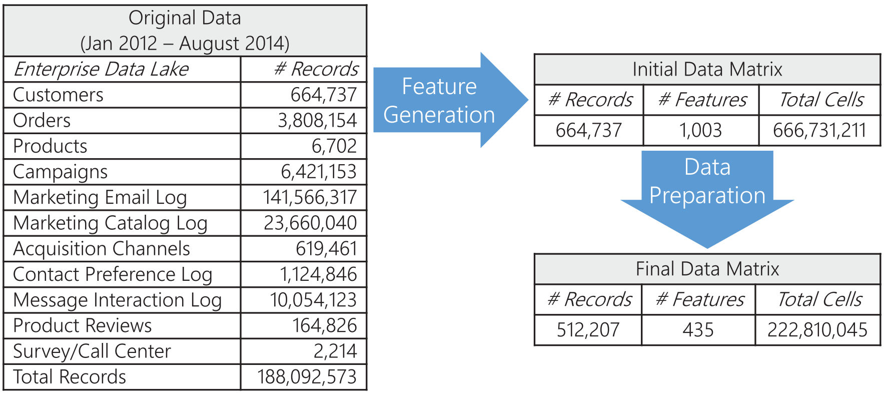

Data Lake
Transaction Support (OLTP)
Customers
Orders
Products
Customer Interaction (CRM)
Campaigns
Marketing Message Log
Acquisition Channels
Contact Preference Log
Message Interaction Log
Voice of Customer
Product Reviews
Surveys/ Call Center
Feature Generation
Product Taxonomy Construction
Association Rule Mining
Engagement Measurement
Satisfaction Approximation
Demographics
Transactions
Choice
Messaging
Channel
Engagement
Satisfaction
Data Matrix
Data Preparation
Data Aggregation/ De-duplication
Missing Value Removal
Outlier Removal
Discretization/ Binarization
Feature Type Conversion
Feature Selection: a. Remove Zero Variance b. Feature Weighting
Balanced Sampling
Training/Test/ Validation Data
Offline Modeling/ Evaluation
Modeling
Windowing (Temporal Effect Adjustment)
Machine Learning Algorithm: Composite Convolution Kernel SVM
Intelligent Experimentation
Offline Evaluation
Add-in/Leave-out Analysis
Top k% Churning Evaluation
Projected Cost-Benefit Analysis
Predictive Model
Online Implementation
Online Evaluation
Field Experiments A/B Testing
Short/Middle-Term Cost-Benefit Analysis
Analytics Infrastructure Implementation
Big Data Management
Big Data Computing: Streamlining/ Parallelization
Feature Engineering Pipeline
Model Building Platform
Iterative Process
IT Constructs
Relational Database
Files/Texts
Intermediate Data
Process
Stage (dashed line means not included in our study)

Figure 3. System Diagram for Advanced Customer Analytics Instantiation



Original Data (Jan 2012 – August 2014)	
Enterprise Data Lake	*# Records*
Customers	664,737
Orders	3,808,154
Products	6,702
Campaigns	6,421,153
Marketing Email Log	141,566,317
Marketing Catalog Log	23,660,040
Acquisition Channels	619,461
Contact Preference Log	1,124,846
Message Interaction Log	10,054,123
Product Reviews	164,826
Survey/Call Center	2,214
Total Records	188,092,573

Initial Data Matrix		
# Records	*# Features*	*Total Cells*
664,737	1,003	666,731,211

Final Data Matrix		
# Records	*# Features*	*Total Cells*
512,207	435	222,810,045

Figure 4. Summary Statistics of Data Related to Customer Churn Analytics

Demographic features included age, gender, income, net worth, education, and household information. With respect to relationship-oriented constructs, *channel* variables were also relatively straightforward to operationalize. These features describe the channels through which a customer was acquired and/or made an initial purchase, including acquisition e-mails, call center upsells, paid search, partners, prospect mailings, radio, social media, and television. However, other relationship-oriented constructs such as *choice, messaging, engagement*, and *satisfaction* necessitated the use of more involved logic and algorithms applied to multiple data sources (as indicated in Figure 3). We discuss these construct categories in the remainder of the section.

In a novel operationalization, our *choice* features focus on the interplay between what products (and categories of products) a customer has purchased, relative to what the company is offering and promoting. Figure 5 illustrates how the choice variables were operationalized using a novel product taxonomy construction to develop a tree of product categories, subcategories, products, and promotions related to products purchased by the customer. Whereas prior studies have included only the specific category of purchase as a variable, our taxonomic representation facilitates richer contextualization of customer choice in the relationship—enabling enhanced discriminatory potential. In the example, the purchased course "The Addictive Brain" is used to generate features such as the number of choices in the same category, subcategory, and promotions bundles received by the customer.

Engagement variables that are related to how a customer interacts with a company provide essential intermediate cues regarding the status of the relationship [19]. However, inclusion in prior studies has typically been limited to service usage in industries such as telecommunications. In the context of CSI, e-mail and clickstreams were prominent avenues for customer engagement. Using the message interaction logs in our data lake, we developed variables related to various engagement actions, including opening, forwarding, and clicking e-mails, as well as viewing and engaging with the landing pages to

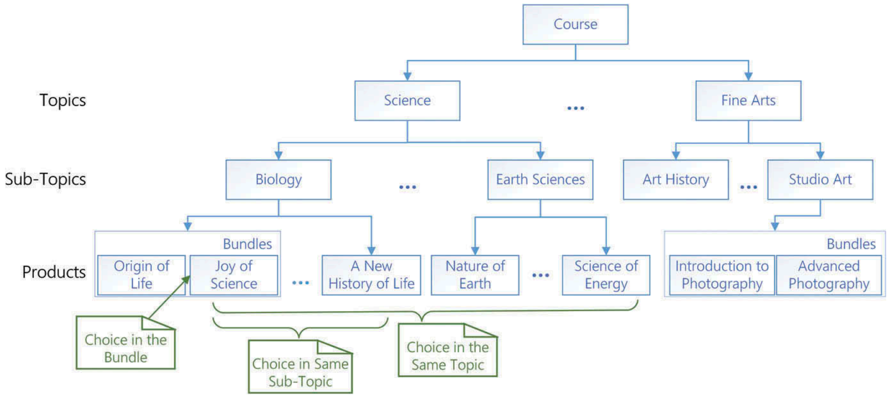


Figure 5. Product Taxonomy and Choice Variables

which they lead. Utilizing contact preference logs, we also developed engagement variables related to customer's current contact preferences, as well as changes in preferences.

Satisfaction features focus primarily on online product reviews provided by customers. CSI currently has no method for linking review authors to their customer database, so in order to incorporate these data, a satisfaction approximation method was employed. Review satisfactions (e.g., ratings, votes) and percentage change in these measures were aggregated at the product level for products purchased by each customer during the first 30 days of initial purchase. These contemporary review characteristics for products purchased by the customer were used as proxy indicators of satisfaction. As later demonstrated, this approach for overcoming the satisfaction integration issue provided significant benefit in our system. We also included measures of satisfaction and other perceptions from surveys and call center logs.

A firm's *messaging* to customers can have a profound impact on how customers perceive the relationship [17, 21] and on future customer behavior. Customers receive a diverse set of mass and customized physical and digital messaging. Given that the mode, quantity, and combination of messaging can impact customer behavior [59], we propose a novel approach for extracting messaging patterns most likely to result in future customer purchases. Adapting highly efficient new methods [65], we employed association rule mining to generate messaging features from hundreds of millions of messaging records. We generated frequent item sets of messages, retaining only those sets that culminated in a purchase in order to find messages strongly associated with purchases. Details regarding this method are provided in online Appendix C. In addition to features related to these association rule-mined messages, we included various other messaging related features, including frequency of mail/e-mail, discounts, promoted products, overlap with categories from a customer's initial purchase, and so forth.

Data Preparation: Once the feature generation stage was completed, an initial data matrix encompassing 1,003 variables for each of the 664,737 customers was constructed, comprising approximately 667 million values. Various necessary data preparation steps were undertaken to handle veracity issues. Features or customer

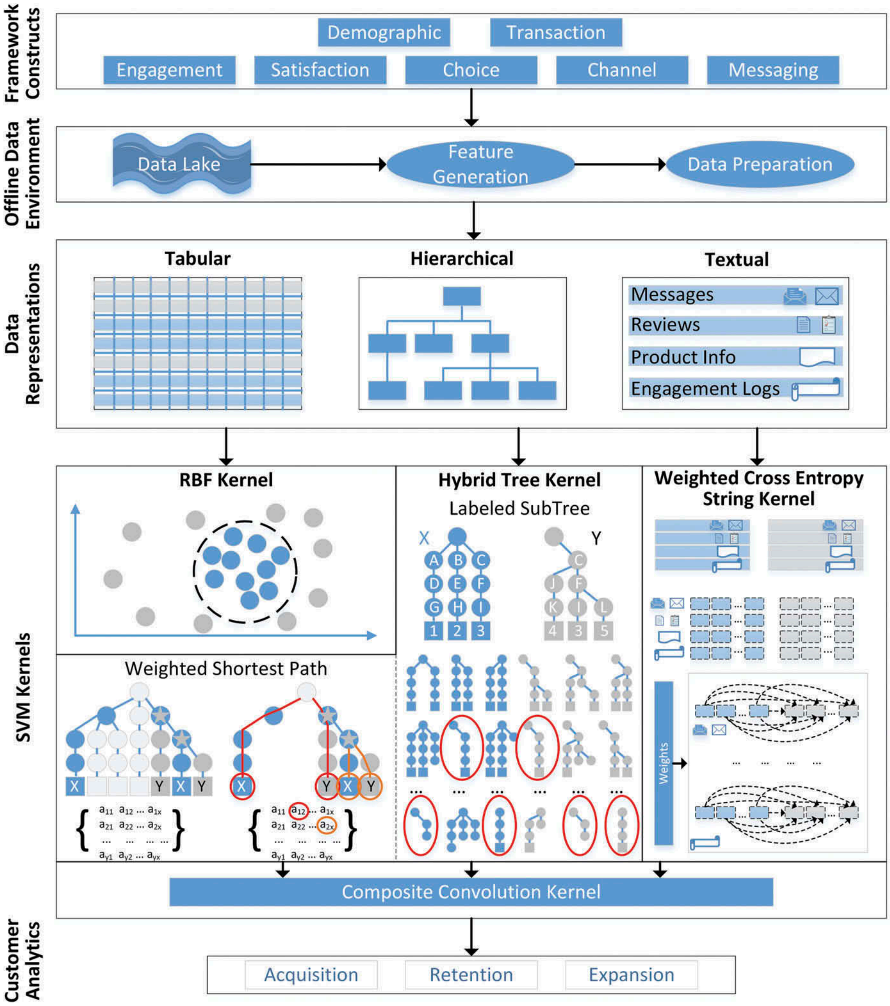


Figure 6. SVM Kernel-Based Customer Analytics Method

records with a high volume of missing values were removed. Outlier removal was applied to eliminate records with abnormal entries (e.g., negative age or no transaction history). Feature weighting by information gain and chi-squared statistics was used to identify features with zero variance or negligible information. Ultimately, 435 features were retained. Moreover, random undersampling of the majority class (i.e., churn) was used on the training data to achieve class balance.

Model

Offline Modeling/Evaluation: Kernel-based machine learning methods have garnered attention from the information systems (IS) community in recent years for their ability to derive patterns from large quantities of heterogenous, noisy, high-velocity data [3, 4]. These methods have outperformed state-of-the-art rule-based, tree-based, Bayesian, and deep learning models in recent benchmarking studies pertaining to voice-of-the-customer tasks [4], while simultaneously providing the added benefit of greater transparency than other big data machine learning methods through greater explanatory potential and provisions for theory-driven design [2]. In sum, kernel-based methods afford the following potential opportunities and benefits for our advanced customer analytics context:

- Principled, theory-driven kernel design by leveraging key customer nuances elucidated by RMT
- Capability to effectively incorporate different types of customer–firm interaction patterns manifested in diverse structured and unstructured enterprise data through use of custom kernels designed for tabular, graphical, and string-based inputs
- Potential to efficiently fuse these diverse custom kernels through a meta-level composite convolution kernel, providing robust and flexible ensemble-like performance capabilities across a wide portfolio of customer analytics tasks

The novel kernel-based SVM customer analytics method we propose is depicted in Figure 6. We begin with the RMT-based constructs previously described, including engagement, satisfaction, choice, channel, and messaging. Next we leverage the aforementioned enterprise customer analytics system encompassing an extensive *offline data environment* coupled with an online real-time processing module (see Figure 3 for details). In particular, the data lake, feature generation, and data preparation modules of the system are used to convert raw customer relationship inputs into features and/or formats suitable for our SVM composite convolution kernel. This results in three distinct data representations. The tabular representation is composed of the 435 features detailed in Figure 4. and Figure 4. The hierarchical representation utilizes the product portfolio taxonomy (Figure 5), both in context of customer purchases and promotional offers. The *textual* data incorporate four types of strings:

- *Messages* include all text of online/e-mail-based marketing initiated by the firm
- *Engagement Logs* include string representations of select customer interactions with the firm, including mail/e-mail preference changes and open/click/view activities
- *Reviews* include title and body text from reviews of products purchased by the customer
- *Product Info* contains the title and description text of purchased products

The tabular, hierarchical, and textual representations feed into the composite convolution kernel, composing three underlying kernels: radial basis function (RBF), hybrid tree (HT), and weighted cross entropy string (WCES). The main intuition guiding our composite convolution setup is that customer–firm relationships embody dynamic, multifaceted patterns that may encompass a plethora of manifestations, including point-value quantifications of consumer decisions and actions, structural representations of latent preferences, and semantic/stylistic indicators of customer proclivities. A critical aspect of kernel-based methods is the kernel matrix comprising the similarity scores between any two training instances. Next we describe each of the underlying kernels at a high level (online Appendix D includes additional details).

Using the tabular data representation, the RBF kernel uses a Gaussian classifier to capture nonlinear "localized learning" patterns from point-value features [10]. The HT kernel combines two tree methods [15]: a novel weighted shortest path tree approach and a subtree method, both utilizing the hierarchical product line taxonomy. The shortest path approach utilizes two large global probabilistic trees encompassing all product purchases for positive and negative class cases observed in the training set (e.g., churn and return). Link strengths between tree nodes are proportional to product co-occurrence likelihood across all class instances in the training set. Any two customers in the training set are compared on both trees by computing the shortest paths between all nodes in their respective trees. To account for potentially nuanced complementary/substitutive relations between product purchases across customers, a purchase co-occurrence matrix is used to weight similarities. In Figure 6, the Weighted Shortest Path illustration depicts an example involving two customers (X and Y nodes), each with a two purchases (leaf nodes are products, nonleaf nodes categories).

The second half of the HT kernel utilizes a labeled subtree method. Whereas the shortest path approach is well-suited to capture probabilistic similarity between customer preferences, subtrees are effective in incorporating structural taxonomic similarity [15] such as commonalities between customer preferences for certain specific categories, or category breadth versus depth similarities. Instead of relying on global trees, the subtree method only compares the customer purchase trees. For instance, in the Labeled SubTree example shown in Figure 6, X and Y gray customers each purchased three items, including one common item (#3). The subtree method compares all unique subtrees from each customers' purchase tree. Similarity between any two customers is computed as the proportion of matching subtrees from their respective purchase trees. In both the shortest path and subtree kernels, we also incorporate various constructs from the tabular representation for each customer order item (i.e., a single block).

WCES is a novel string kernel [37] geared toward uncovering semantic and stylistic customer tendencies hidden in text descriptions or logs. All training texts are tokenized separately within each of the four categories of text, and unigram tokens are weighted using the information gain heuristic. Due to class imbalance, undersampling, and use of a single training set, we also incorporate a feature stability heuristic to as part of our token weights to alleviate overfitting [33]. These weights are used as part of a cross entropy-based customer similarity scoring mechanism. Cross entropy has been shown to be effective in identifying commonalities in unstructured data distributions [30]. In our context, it can help identify hidden commonalities in customer proclivities based on purchase of introductory versus advanced products or certain styles or genres of offerings, specific sticky promotional language, specialized purchase use cases, subtle communication and interaction indicators, and so forth. As depicted in Figure 6, for each customer, we randomly extract a predefined number of text windows of certain length (e.g., 50 characters) for each of the four textual representations. All such window pair combinations between any two customers are compared using a weighted similarity comparison.

At the composite kernel stage, the RBF, HT, and WCES customer similarity kernel matrices are fused using multiple kernel learning to allow robust predictive power across an array of customer acquisition, retention, and expansion-related customer analytics tasks. We employ this composite convolution kernel SVM model within our prototype system to predict customer behaviors across a portfolio of applications. For comparison we also examined more traditional machine learning algorithms, including stochastic gradient boosting, CART, boosted generalized linear model, C5.0, random forest, and naive Bayes. Because of its ability to incorporate a richer set of structured and unstructured data, the custom kernel SVM solution described here significantly outperformed these methods. However, we found that the more important factor for performance was the variety of data constructs provided to the models. For further details, see online Appendix E.

Evaluation Metrics: To evaluate our system as well as determine the value of individual data constructs, we created evaluation metrics based on the value of actions taken as a result of model predictions. For the task of churn prediction, the proposed action to be taken based on the model is to discontinue marketing physical catalog mailings to the top 10 percent of customers most likely to churn. According to CSI, the average cost of catalog mailings over the life of a customer is $66, and the average lifetime revenue from a customer who does not churn after the first purchase is $260. The costs and benefits of system-driven actions for the CLV and conversion tasks are detailed in online Appendix B. Applied across an expected population of 250,000 new customers each year, these figures are used to arrive at the overall value for our system as well as each individual construct category.

Online Implementation: After building the prototype system, we delivered the system and our results to CSI. They are currently testing to validate expected outcomes for customers and creating a plan to implement the system in a production capacity. In addition, based on the results of construct evaluation discussed next, CSI has already made investments in infrastructure for data management and integration (discussed at the end of the following section), which provides support for the core benefit of our framework

Evaluation

Through evaluation of our prototype system implementation, we demonstrate the utility as well as economic value of our contributed artifacts [49]. In the setting of CSI, where customer outcomes are highly uncertain and costs of serving customers are high, early predictions of customer behaviors are critical. Therefore, we use only customer characteristics observable within the first 30 days of an initial purchase as inputs to system predictions.[2] To ensure realistic results, we utilized chronological evaluation with nine-month rolling windows for training, and one-month windows for capturing test observations.

Table 1. Add-In and Leave-Out Construct Evaluation

	Accuracy for top 10% most likely to churn	Annualized net cost savings, $	Net savings as % of total marginal marketing expense	Difference from baseline model (z-score)	Difference from full model (z-score)
Baseline Model	54.56	–1,598,979	–9.69	–	61.26***
Add-In Channel	57.22	–1,130,701	–6.85	3.82***	57.92***
Add-In Satisfaction	67.84	–440,311	–2.67	19.45***	44.01***
Add-In Engagement	69.73	–317,802	–1.93	22.32***	41.39***
Add-In Choice	71.44	–206,140	–1.25	24.96***	38.96***
Add-In Messaging	78.46	250,079	1.52	36.15***	28.29***
Leave-Out Messaging	89.74	983,219	5.96	56.01***	6.76***
Leave-Out Satisfaction	90.02	1,001,085	6.07	56.54***	6.12***
Leave-Out Engagement	90.67	1,043,835	6.33	57.80***	4.54***
Leave-Out Choice	91.74	1,113,385	6.75	59.88***	1.84
Leave-Out Channel	92.17	1,140,822	6.91	60.71***	0.74
Full Model	92.44	1,158,688	7.02	61.26***	—

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Construct Evaluation: Customer Retention–Churn Application

Key features of our proposed framework for advanced customer analytics include: (1) the comprehensive set of constructs to guide identification and acquisition of data for advanced customer analytics applications; and (2) the mechanism for identifying the value of these various constructs. By measuring the value of each individual construct, IT managers asked to support big data analytics initiatives can make better-informed decisions regarding investment in data management, integration, and real-time feature engineering, and more effectively justify efforts in these endeavors. To accomplish this, we trained models using (1) all construct categories,[3] (2) only transaction and demographics as a baseline, (3) each individual category left out of the full model, and (4) each individual category added in to the baseline. The results of these models are presented in Table 1. The full model provides 92.44 percent accuracy for the top 10 percent of customers most likely to churn, which would provide a net cost savings of 7.0 percent of CSI's total marginal marketing spend. The baseline model, on the other hand, has an accuracy of only 54.56 percent for the top 10 percent, resulting in a net loss of revenue of 9.7 percent of total marginal marketing spend.

Table 1 reports z-scores for tests of differences in models' top 10 percent accuracies. Comparing add-in models to the baseline, each construct added significantly improves the model accuracy. Engagement, choice, and messaging appear to be the most valuable constructs from the add-in perspective, improving accuracy significantly over the add-in satisfaction model, which, in turn, provides a significant improvement over the add-in channel. The add-in comparisons are useful for providing initial direction to determine which single construct could add the most value if focused on first. From the leave-out perspective, engagement, satisfaction, and messaging are the most important constructs. Depending on costs of data management, integration, and real-time feature engineering of various constructs, the model could be further tested with various construct subsets to determine the set providing optimal value. It is important to note that all of the leave-out models significantly outperform all of the add-in models, pointing to significant synergy among the relationship-oriented constructs in making accurate predictions.

System Evaluation: Customer Retention–Churn Application

In addition to evaluating the set of constructs proposed and demonstrating the use of our framework to search for a solution that provides optimal strategic value to the organization, we compared our prototype system to several existing models to demonstrate its value. As mentioned previously, no research has focused on models that would be effective in high single-purchase, low frequency environments, which made the search for benchmark models challenging. Almost all models created for churn explanation and prediction rely on RFM constructs [20, 52, 55] that provide little variation with which to predict in this environment, or perceptual constructs that are difficult or impossible to measure for purposes of individual-level prediction [7, 17]. Also as mentioned, much of the research examining churn prediction is focused solely on innovation or choice in modeling techniques, with no regard to the constructs or general framework for designing the analytics [36, 60]. While this research is valuable, it does not provide a valid benchmark for our prototype system, as our system is composed of not only a novel method but also the set of relationship-oriented data constructs supporting the method. For benchmark comparison, we had to find approaches that encompassed a framework of constructs for prediction as well as a model. Taking into account each of these challenges, we extensively surveyed the literature to identify the most advanced, applicable approaches available.

The first benchmark identified was that of the classic RFM model. We implemented the beta geometric/negative binomial distribution (BG/NBD) model of Fader et al. [20].[4] This model is considered the standard for churn prediction, and it, or a less sophisticated analogue, is widely used by many organizations for churn prediction (including CSI, prior to this work). Next, we identified other models that also used RFM constructs, but added additional features for prediction. Buckinx and Van den Poel [9] develop a churn prediction system that relies on RFM constructs, other transaction characteristics, customer choices, promotional messaging, and demographics and leverages logistic regression, random forest, and neural network models. Coussement and De Bock [16] utilize general additive models, decision trees, and random forests to predict churn using RFM and

Table 2. Comparison to Benchmark Models

	Model type	Accuracy for top 10% most likely to churn	Annualized net cost savings, $	Net savings as % of total marginal marketing expense
Prototype system	**Composite convolution kernel SVM**	**92.4**	**1,158,688**	**7.0**
Ballings and Van den Poel [6]	Logistic regression	86.1***	747,636	4.5
	Decision tree	81.8***	464,351	2.8
	Bagged decision trees	83.4***	568,719	3.4
Buckinx and Van den Poel [9]	Logistic regression	86.8***	795,109	4.8
	Random forest	83.4***	570,460	3.5
	Neural network	82.7***	522,336	3.2
Chen and Hitt [14]	Logistic regression	86.9***	797,035	4.8
Coussement and De Bock [16]	Generalized additive models	86.2***	755,287	4.6
	Decision tree	83.1***	549,650	3.3
	Random forest	83.9***	605,577	3.7
Fader et al. [20]	BG/NBD	81.9***	471,847	2.9
Mittal and Kamakura [44]	Probit binary choice	85.5***	707,757	4.3
Neslin et al. [45]	Logistic regression	84.4***	634,536	3.8
Vanderveld and Han [58]	Random forest	85.4***	703,126	4.3

***Model performance inferior to prototype system at significance of $p < .001$.

demographic characteristics. Neslin et al. [45] formulate an approach using logistic regression with RFM and other transaction features along with messaging characteristics to predict churn, specifically focusing on overcoming the "recency" trap, wherein many customers have not purchased recently. Ballings and Van den Poel [6] use decision trees and logistic regression to predict churn using RFM and other transaction features along with demographics. We also identified models that used more perceptual constructs in the tradition of Crosby and Stephens [17]. Specifically, Mittal and Kamakura [44] employ measures of satisfaction combined with demographic characteristics to predict churn using a probit choice model. Chen and Hitt [14] predict churn and switching behavior with a logistic regression model including measures of satisfaction and engagement through website usage characteristics as well as demographic information. Finally, Vanderveld and Han [58] focus on engagement features to predict churn using random forests.

Each of these models employs a different prediction algorithm and set of constructs. In order to compare them to our prototype system, we operationalized each construct

Table 3. Customer Analytics Portfolio Value, $

	Value for Task 1: Churn	Value for Task 2: CLV	Value for Task 3: Conversion	Total Value
Baseline Model	N/A*	578,380	N/A*	578,380
Add-in Channel	N/A*	715,453	N/A*	715,453
Add-in Engagement	N/A*	929,421	N/A*	929,421
Add-in Satisfaction	N/A*	1,053,120	N/A*	1,053,120
Add-in Choice	N/A*	1,068,722	N/A*	1,068,722
Add-in Messaging	250,079	1,359,584	1,624,833	1,359,584
Leave-out Messaging	983,219	1,462,110	744,645	3,189,974
Leave-out Choice	1,113,385	1,739,599	1,717,484	4,570,468
Leave-out Satisfaction	1,001,085	1,562,407	2,180,741	4,744,233
Leave-out Engagement	1,043,835	1,584,695	2,366,044	4,994,575
Leave-out Channel	1,140,822	1,770,802	2,597,672	5,509,296
Full Model	1,158,688	1,784,175	2,759,812	5,702,675

*Model produces negative value; hence it would be foregone and therefore is not included in the total value calculation.

included in their models based on data available for CSI. For instance, satisfaction constructs were operationalized through online reviews contemporary to a customer's initial purchase, and engagement was operationalized through e-mail open and click rates, as well as opt-in or opt-out of communications from CSI, as described previously. Each model was evaluated using the same windowing strategy as the prototype system.

As shown in Table 2, the prototype system outperforms each of the benchmark models by a wide margin. Based on *z*-tests for differences in model accuracies for the top 10 percent most likely churners, the prototype system has significantly higher accuracies than all other models at a significance of $p < .001$. The improvement of the prototype over existing models is also highly economically significant, with net savings of 7.0 percent of total marginal marketing spending, as compared to 2.9 percent to 4.8 percent saved by other models. This is despite the fact that the benchmark models comprise leading approaches from marketing, IS, and machine learning disciplines.

Construct and System Evaluation: Portfolio

Our prototype system includes a portfolio of three customer analytics applications:

- A churn prediction application focused on evaluating which customers the firm should invest in through continued marketing efforts (retention)
- A conversion prediction application for identifying customers likely to respond to individual e-mail promotions to reduce messaging fatigue and prevent attrition (retention/expansion)

- A CLV prediction application for identifying customers who could successfully be expanded through offer of participation in a premium loyalty program (expansion)

The CLV and conversion applications are detailed in online Appendix B. Table 3 demonstrates the total value achieved across the analytics portfolio as a whole, given the inclusion or exclusion of various construct categories. Along with satisfaction and engagement, which provided value to the churn prediction model, choice and messaging are shown to have significant value across the portfolio of analytics applications. This analysis illustrates how our proposed framework may be used to develop a platform for agility in deploying customer analytics to achieve strategic value through big data.

If, for example, investments in infrastructure were made to integrate all available constructs into a feature generation pipeline for live deployment of the full model, the estimated annualized value provided by the analytics portfolio would be $5,702,675. This value (and that of future analytics) may be compared with the costs of the infrastructure investment as a whole, as well as for each data construct/source. For instance, if it is determined that the cost of obtaining and integrating data for the engagement construct exceeds $708,100 (per annum), the firm may choose to implement a system without it, resulting in a reduced $4,994,575 annualized value. In addition to the various options shown in Table 3, other construct subsets may be tested to choose the best complement of data. This evaluation provides value-based justification for IT investment in data management and integration efforts to support analytics initiatives.

Results Discussion

The instantiation of a portfolio of advanced customer analytics applications demonstrates how our framework may be leveraged to develop capabilities for agility through big data analytics and create strategic value and sustainable competitive advantage in a dynamic market environment. By creating a value-justified infrastructure for data integration and management to support advanced customer analytics, firms can create a portfolio of analytics applications to serve a variety of strategic purposes, providing significant value. The applications in our portfolio combine to generate nearly $6 million annualized value and represent only a small fraction of the opportunity for deploying analytics from this infrastructure to strengthen and leverage customer relationships.

Evaluation of the system demonstrates that advanced customer analytics systems built on relationship-oriented data from a variety of sources can accurately predict customer behavior and add value. The prototype system for churn prediction significantly outperformed each of the leading benchmarks for comparison, including the well-adopted BG/NBD approach, evidencing the impetus for relationship-oriented advanced customer analytics supported by IT-enabled data infrastructure. Further, in addition to our system based on the composite convolution kernel SVM, we also evaluated a system using more traditional machine learning techniques (see online Appendix E), and found that the SVM provides a significant lift in performance over the best alternative model, C5.0 (92.4 percent vs. 91.0 percent accuracy and $5.7 vs. $5.1 million net benefit), but variations in input data had a larger impact than model choice.

Our results also demonstrate how, through use of the proposed framework, the individual value of various data sources may be identified for use in prioritization of data management and integration efforts. Each of the construct categories identified aside from channel was shown to add significant value over the portfolio of analytics applications. Engagement provides significant value across all models, as suggested during construct justification because of the information it provides regarding customers' continuing interest despite a lack of transactions. When left out of the portfolio, it reduces value provided by the analytics initiative by an estimated $708,100 per annum. Messaging and choice, with rich information about customer preferences and communications, each provide even more value ($2,512,701 and $1,132,207, respectively). Satisfaction is also found to be important to the portfolio, which may be expected due to its near ubiquitous prominence in the RMT literature [7, 17, 44]. The $958,442 value is provided despite the fact that operationalizations rely on contemporary review data to impute satisfaction levels for individual customers rather than surveying every individual customer. There are also significant synergies from the use of multiple data sources, which may be seen by the value of the messaging construct. Alone it provides less than $800,000 in value over the baseline, but its combination with other constructs contributes over $2.5 million to the full model. This stresses the importance of combining rich relationship-oriented data across silos, as well as the need for a data valuation framework.

These values should be compared with data collection and integration costs in order to guide investments in data infrastructure. The strength of results regarding engagement and satisfaction constructs despite their limited available operationalizations have already led CSI to invest in changes to infrastructure to support data management and integration. First, CSI has invested in a new platform for managing its web presence in order to support the identification of individual customers browsing its site and integration of these data into existing customer management software. Prior to this work, CSI could analyze web traffic, but not tie it to individual customers for performing analytics. Second, CSI has invested in infrastructure that will allow it to tie online product reviews to the individual customers posting them. Both of these changes were implemented as a direct consequence of the results of this study, which provided IT management with the support and justification to make these investments in data management and integration to support its advanced customer analytics initiatives.

Conclusion

In this study, we present a framework for designing advanced customer analytics solutions based on relationship-oriented constructs. This work encompasses three key contributions. First and foremost, we contribute to the design science literature in the creation of a synergistic ecosystem of novel IT artifacts for performing advanced customer analytics in the era of big data [28]. Our framework provides guidance for the agile development and deployment of advanced customer analytics solutions that predict customer behavior and inform strategic business decisions. Following guiding principles from our framework, we develop a novel kernel-based machine learning method that is custom-designed to extract insight and value from a rich variety of relationship-oriented data constructs. We also provide a prototype system instantiation of a portfolio of advanced customer analytics applications for a firm with high proportions of single or infrequent purchase customers, a problem that cannot be addressed by the siloed approaches of existing customer analytics methods [6, 41]. The system represents a rigorous proof-of-concept, while its results offer practical relevance [28]. We show that this system enables significant strategic value, contributing nearly $6 million in estimated annualized benefits. This value will increase with additional analytics supported by the value-justified data infrastructure informed by our framework.

Second, we contribute to managerial practice for firms attempting to employ big data analytics to drive strategic value. Organizations are overwhelmed by available data [40], and IT managers who are asked to support big data analytics through data management and integration are in need of a blueprint for valuing various data sources and justifying their efforts through return on investment in infrastructure [11, 35]. The framework provides a structure through which the various relationship-oriented constructs can be evaluated based on added business value relative to costs of data acquisition, management, integration, and real-time feature construction. Our contributions in this area are validated by the direct impact of our results for CSI in motivating strategic investment in data management and integration infrastructure.

Third, our research contributes to the nascent literature regarding predictive analytics through the use of big data [2, 5, 13, 22]. The kernel theory that we employ in our design, RMT, has investigated many of the constructs that are central to our framework [7, 17, 44]. However, all of this work has been from an explanatory, rather than predictive, viewpoint. We provide a firm foundation that allows us to answer recent calls by many in the IS field for predictive analytics [56], particularly utilizing volume and variety of available data [13, 22] to predict micro-level outcomes at the individual level [5, 22]. The advanced customer analytics our framework allows are not feasible in the absence of big data from a variety of sources providing a relationship-oriented view of customer behavior.

In this era of profound digital transformation, customer agility lies at the intersection of customer analytics, big data, and IT strategy. Firms capable of taking advantage of such agility are best positioned to achieve sustainable competitive advantage. Our study makes important contributions to the nascent literature on this critical topic.

Supplemental File

Supplemental data for this article can be accessed on the publisher's website at https://doi.org/10.1080/07421222.2018.1451957

Notes

1. All studies cited in the references for the main body of our manuscript were reviewed, as well as additional papers listed in online Appendix F (over 200 relevant studies in total). While it is not feasible to review all relevant studies, this creates a representative set from which we may draw conclusions about the completeness of our construct set.

2. We evaluated other observation period lengths (60 and 90 days), but the small improvements in predictive power provided were outweighed by the reduction in value caused by longer lead times to predictions.

3. No data for the firm and environment characteristics construct is included in our prototype system. As noted in the description of this construct within the broader framework, although potentially useful for prediction, this construct is likely more informative of shifts suggesting models be revisited as firms monitor analytics solutions after deployment.

4. We would like to thank Peter Fader for his support as we implemented this model, as well as for providing the data set used in his paper so that we could verify our implementation to be identical.

References

1. Abbasi, A.; Albrecht, C.; Vance, A.; and Hansen, J. Metafraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, *36*, 4 (2012), 1293–1327.

2. Abbasi, A.; Sarker, S.; and Chiang, R.H.L. Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association of Information Systems*, *17*, 2 (2016), 1–32.

3. Abbasi, A.; Zahedi, F.; Zeng, D.; Chen, Y.; Chen, H.; and Nunamaker, J.F. Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems*, *31*, 4 (2015), 109–157.

4. Abbasi, A.; Zhou, Y.; Deng, S.; and Zhang, P. Text analytics for sense-making in social media: A language-action perspective. *MIS Quarterly*, forthcoming.

5. Agarwal, R., and Dhar, V. Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *25*, 3 (2014), 443–448.

6. Ballings, M., and Van Den Poel, D. Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications*, *39*, 18 (2012), 13517–13522.

7. Bolton, R.N. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science*, *17*, 1 (1998), 45–65.

8. Bolton, R.N.; Lemon, K.N.; and Verhoef, P.C. The theoretical underpinnings of customer asset management: A framework and propositions for future research. *Journal of the Academy of Marketing Science*, *32*, 3 (2004), 271–292.

9. Buckinx, W., and Van den Poel, D. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*, 1 (2005), 252–268.

10. Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 2 (1998), 121–167.

11. Cappiello, C.; Francalanci, C.; and Pernici, B. Time-related factors of data quality in information multichannel systems. *Journal of Management Information Systems*, *20*, 3 (2003), 71–91.

12. Chen, D.Q.; Preston, D.S.; and Swink, M. How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems*, *32*, 4 (2015), 4–39.

13. Chen, H., and Storey, V.C. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*, 4 (2012), 1165–1188.

14. Chen, P.S., and Hitt, L.M. Measuring switching costs and the determinants of customer retention in internet-enabled businesses: A study of the online brokerage industry. *Information Systems Research*, *13*, 3 (2002), 255–74.

15. Collins, M., and Duffy, N. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*, (2002), Vancouver, British Columbia, Canada. pp. 625–632.

16. Coussement, K., and De Bock, K.W. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*, 9 (2013), 1629–1636.

17. Crosby, L.A., and Stephens, N. Effects of relationship marketing on satisfaction, retention, and prices in the life insurance industry. *Journal of Marketing Research*, *24*, 4 (1987), 404–411.

18. Davenport, T.H. Analytics 3.0. *Harvard Business Review*, December 2013, 64–72.

19. van Doorn, J.; Lemon, K.N.N.; Mittal, V.; et al. Customer engagement behavior: Theoretical foundations and research directions. *Journal of Service Research*, *13*, 3 (2010), 253–266.

20. Fader, P.S.; Hardie, B.G.S.; and Lee, K.L. Counting your customers the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, *24*, 2 (2005), 275–284.

21. Gázquez-Abad, J.C.; Canniére, M.H. De; and Martínez-López, F.J. Dynamics of customer response to promotional and relational direct mailings from an apparel retailer: The moderating role of relationship strength. *Journal of Retailing*, *87*, 2 (2011), 166–181.

22. Goes, P.B. Big data and IS research. *MIS Quarterly*, *38*, 3 (2014), iii–viii.

23. Goodhue, D.L.; Kirsch, L.J.; Quillard, J.A; and Wybo, M.D. Strategic data planning: Lessons from the field. *MIS Quarterly*, *16*, 1 (1992), 11–34.

24. Goodhue, D.L., Wybo, M.D., and Kirsch, L.J. The impact of data integration on the costs and benefits of information systems. *MIS Quarterly*, *16*, 3 (1992), 293–311.

25. Gunarathne, P.; Rui, H.; and Seidmann, A. Whose and what social media complaints have happier resolutions? Evidence from Twitter. *Journal of Management Information Systems*, *34*, 2 (2017), 314–340.

26. Gupta, S.; Hanssens, D.; Hardie, B.; et al. Modeling customer lifetime value. *Journal of Service Research*, *9*, 2 (2006), 139–155.

27. Heudecker, N., and White, A. The data lake fallacy: All water and little substance. *Gartner*, July 2014, 6.

28. Hevner, A.R.; March, S.T.; Park, J.; and Ram, S. Design science in information systems research. *MIS Quarterly*, *28*, 1 (2004), 75–105.

29. Hitt, L.M., and Frei, F.X. Do Better customers utilize electronic distribution channels? The case of PC banking. *Management Science*, *48*, 6 (2002), 732–748.

30. Juola, P., and Baayen, H. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing, 20* (2005), 59–67.

31. Karimi, J., and Walter, Z. The role of dynamic capabilities in responding to digital disruption: A factor-based study of the newspaper industry. *Journal of Management Information Systems*, *32*, 1 (2015), 39–81.

32. Keane, T.J., and Wang, P. Applications for the lifetime value model in modern newspaper publishing. *Journal of Direct Marketing*, *9*, 2 (1995), 59–66.

33. Koppel, M.; Akiva, N.; and Dagan, I. Feature instability as a criterion for selecting potential style markers. *JASIST*, *57*, 11 (2006), 1519–1525.

34. Kunz, W.; Aksoy, L.; Bart, Y.; et al. Customer engagement in a big data world. *Journal of Services Marketing*, *31*, 2 (2017), 161–171.

35. Laney, D. Why and how to measure the value of your information assets. *Gartner*, August 2015. 1–22

36. Lemmens, A., and Croux, C. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*, 2 (2006), 276–286.

37. Lodhi, H.; Saunders, C.; Shawe-Taylor, J.; Cristianini, N.; and Watkins, C. Text classification using string kernels. *Journal of Machine Learning Research, 2* (2002), 419–444.

38. Lu, Y., and Ramamurthy, K. Understanding the link between information technology capability and organizational agility: An empirical examination. *MIS Quarterly*, *35*, 4 (2011), 931–954.

39. Lyytinen, K., and Grover, V. Management misinformation systems: A time to revisit? *Journal of the Association for Information Systems*, *18*, 3 (2017), 206–230.

40. McAfee, A., and Brynjofsson, E. Big data: The management revolution. *Harvard Business Review*, October 2012, 60–68.

41. Miglautsch, J. Application of RFM principles: What to do with 1–1–1 customers? *Journal of Database Marketing*, *9*, 4 (2002), 319–324.

42. Miguéis, V.L.; Van den Poel, D.; Camanho, A.S.; and Falcão e Cunha, J. Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, *39*, 12 (2012), 11250–11256.

43. Mithas, S.; Ramasubbu, N.; and Sambamurthy, V. How information management capability influences firm performance. *MIS Quarterly*, *35*, 1 (2011), 237–256.

44. Mittal, V., and Kamakura, W.A. Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of Marketing Research, 38* (February 2001)., 131–142.

45. Neslin, S.A.; Taylor, G.A.; Grantham, K.D.; and McNeil, K.R. Overcoming the "recency trap" in customer relationship management. *Journal of the Academy of Marketing Science*, *41*, 3 (2013), 320–337.

46. Neslin, S.A; Gupta, S.; Kamakura, W.; Lu, J.; and Mason, C.H. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43* (May 2006), 204–211.

47. Nunamaker, J.F.; Briggs, R.O.; Derrick, D.C.; and Schwabe, G. The last research mile: Achieving both rigor and relevance in information systems research. *Journal of Management Information Systems*, *32*, 3 (2015), 10–47.

48. Palmer, A. The evolution of an idea: An environmental explanation of relationship marketing. *Journal of Relationship Marketing*, *1*, 1 (2002), 79–94.

49. Prat, N.; Comyn-Wattiau, I.; and Akoka, J. A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, *32*, 3 (2015), 229–267.

50. Ransbotham, B.S., and Kiron, D. Analytics as a source of business innovation. *MIT Sloan Management Review*, February 2017, 1–16.

51. Ransbotham, S.; Kiron, D.; and Prentice, P.K. Minding the analytics gap. *MIT Sloan Management Review, 56*, (Spring 2015), 63–68.

52. Reinartz, W.J., and Kumar, V. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing, 67* (January 2003), 77–99.

53. Roberts, N., and Grover, V. Leveraging information technology infrastructure to facilitate a firm's customer agility and competitive activity: An empirical investigation. *Journal of Management Information Systems*, *28*, 4 (2012), 231–270.

54. Sambamurthy, V.; Bharadwaj, A.; and Grover, V. Shaping agility through digital options: Reconceptualizing the role of information technology in contemporary firms. *MIS Quarterly*, *27*, 2 (2003), 237–263.

55. Schmittlein, D.C.; Morrison, D.G.; and Colombo, R. Counting your customers: Who are they and what will they do next? *Management Science*, *33*, 1 (1987), 1–24.

56. Shmueli, G., and Koppius, O.R. Predictive analytics in information systems research. *MIS Quarterly*, *35*, 3 (2011), 553–572.

57. Teo, T.S.H., and King, W.R. Integration between business planning and information systems planning: An evolutionary-contingency perspective. *Journal of Management Information Systems*, *14*, 1 (1997), 185–214.

58. Vanderveld, A., and Han, A. An engagement-based customer lifetime value system for e-commerce. In *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2016.

59. Venkatesan, R., and Kumar, V. Framework for customer selection. *Journal of Marketing, 68* (October 2004), 106–125.

60. Verbraken, T.; Verbeke, W.; and Baesens, B. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 5 (2013), 961–973.

61. Verhoef, P.C. Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing, 67* (October 2003).

62. Voss, G.B.; Godfrey, A.; and Seiders, K. How complementarity and substitution alter the customer satisfaction–repurchase link. *Journal of Marketing, 74* (November 2010), 111–127.

63. Wagner, C., and Majchrzak, A. Enabling customer-centricity using wikis and the wiki way. *Journal of Management Information Systems*, *23*, 3 (2007), 17–43.

64. Walls, J.G.; Widmeyer, G.R.; and Sawy, O.A. El. Building an information system design theory for vigilant EIS. *Information Systems Research*, *3*, 1 (1992), 36–59.

65. Wang, K., and Skadron, K. Association rule mining with the Micron Automata Processor. *IEEEInternational Parallel and Distributed Processing Symposium*Hyderabad, India. pp. 689–699, 2015.

66. Wixom, B., and Ross, J. How to monetize your data. *MIT Sloan Management Review*, January 2017, 10–13.