



## **A Comparative Analysis of Data Mining Techniques for Customer Relationship Management**

Fawzia El-Masry

*Department of Technology and Innovation Management, Visayan Maritime Polytechnic University, Philippines*

*fawzia.el.masry@vmpu-ph.net*

### **Peer Review Information**

*Submission: 06 Jan 2017*

*Revision: 22 Feb 2017*

*Acceptance: 09 March 2017*

### **Keywords**

*Customer Relationship Management, Data Mining, Customer Churn, Customer Segmentation, Predictive Analytics*

### **Abstract**

Customer relationship management (CRM) depends on the ability to convert large volumes of customer, transaction, service, and interaction data into decisions that improve acquisition, retention, segmentation, and customer value. This methodology paper presents a comparative framework for evaluating widely used data mining techniques in CRM. The study considers descriptive techniques such as clustering and association-rule mining together with predictive techniques including logistic regression, decision trees, random forests, k-nearest neighbors, support vector machines, and neural-network-oriented approaches reported in prior literature. The proposed methodology integrates CRM data preparation, feature engineering, stratified model validation, and multi-metric comparison using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). A fixed-seed synthetic CRM benchmark is used only to demonstrate the evaluation procedure without claiming access to proprietary customer records. The benchmark shows why algorithm selection must reflect the CRM objective: high overall accuracy can conceal weak churn recall, while interpretable statistical models may remain competitive after appropriate data preparation. The paper concludes that CRM analytics should combine predictive performance, interpretability, class-imbalance handling, scalability, and managerial actionability rather than selecting a technique by a single accuracy value.

### **Introduction**

Customer relationship management has evolved from a database-supported marketing function into a decision framework in which firms integrate customer information to support acquisition, development, retention, and service. Early work on marketing knowledge management showed that data mining can transform large customer databases into actionable decision support rather than merely historical reporting (Shaw et al., 2001). From a machine-learning perspective, business data mining also requires matching the analytical method to the business objective, data structure, and cost of errors (Bose and Mahapatra, 2001).

These ideas remain central to CRM because customer data are heterogeneous, behavioral patterns change over time, and different managerial questions require different analytical functions.

Rygielski et al. (2002) described CRM data mining through classification, regression, clustering, association analysis, time-series analysis, and sequence discovery. In practice, these techniques serve distinct purposes. Clustering supports customer segmentation, association rules reveal product or behavior affinities, and classification models estimate outcomes such as churn, campaign response, or

credit-related behavior. Ngai et al. (2009), after reviewing CRM data-mining research, found that customer retention had received substantial attention and that classification and association approaches were among the most frequently applied. Consequently, a meaningful comparative analysis should not ask which algorithm is universally best; it should examine how model characteristics align with the CRM task and evaluation criterion.

A persistent challenge is that customer outcomes are often imbalanced. Churners, responders, or high-risk customers may form a minority class, making raw accuracy misleading. Burez and Van den Poel (2009) demonstrated that class imbalance in churn analysis requires more appropriate performance measures such as AUC and lift, while Neslin et al. (2006) showed that methodological choices can materially affect the economic usefulness of churn models. More recent pre-2018 studies similarly emphasize feature engineering, data preparation, and model comparison rather than reliance on one technique (Moro et al., 2014; Coussement et al., 2017).

The objective of this paper is therefore to develop a clear and reproducible methodology for comparing data mining techniques in CRM. The framework evaluates predictive models with common metrics, positions clustering and association mining within their appropriate descriptive roles, and links analytical results to practical CRM decisions. All scholarly references used are from 2017 or earlier, as required.

## Literature Review

Wei and Chiu (2002) demonstrated how telecommunications call-detail behavior could be transformed into churn predictors and evaluated through classification models. Their work highlighted the importance of recent behavioral changes rather than depending only on static demographics. Hwang et al. (2004) connected customer lifetime value, loyalty, and defection probability to segmentation, illustrating that CRM decisions should distinguish current value, potential value, and retention risk. These studies established two core CRM uses of data mining: prediction of customer actions and segmentation of customer value.

In retail settings, Chen et al. (2005) applied association-oriented analysis to detect changes in customer behavior over time. Association rules are attractive because they produce explicit relationships that managers can use for cross-selling and promotion design. However, association measures do not directly provide a probability of churn or response and can generate many redundant rules. For segmentation, clustering methods such as k-means are useful when customer labels do not exist. Murray et al. (2017) further showed that data-mining-based segmentation can recover meaningful behavioral groups even when conventional descriptive variables are limited or noisy.

Predictive CRM research has compared statistical, tree-based, instance-based, and machine-learning approaches. Buckinx and Van den Poel (2005) compared logistic regression, neural networks, and random forests for partial customer defection and used both classification accuracy and AUC. Keramati et al. (2014) compared decision trees, artificial neural networks, k-nearest neighbors, and support vector machines for telecommunications churn. Vafeiadis et al. (2015) also performed a comparative churn study and showed that boosting can improve several base learners. These studies indicate that algorithm performance is data-dependent: nonlinear models can capture interactions, but complexity can increase tuning requirements and reduce interpretability.

Customer-response modeling provides a related CRM context. Moro et al. (2014) used real bank telemarketing data and emphasized feature engineering and realistic validation when comparing data-mining models. By 2017, comparative research had increasingly focused on the full analytical pipeline. Coussement et al. (2017) showed that alternative data-preparation procedures can substantially alter churn-prediction performance and that an optimized logistic model can remain competitive with advanced algorithms. Khodabandehlou and Zivari Rahman (2017) likewise compared supervised learning approaches for customer churn and emphasized behavioral variables, bagging, and boosting. Collectively, the literature suggests that preprocessing, validation design, class balance, and the business loss function are at least as important as the algorithm name.

**Table 1.** Comparison of major data mining techniques for CRM

| Technique | Primary CRM Use | Main Strength | Main Limitation |
|-----------|-----------------|---------------|-----------------|
|-----------|-----------------|---------------|-----------------|

|                     |                                  |                              |                                       |
|---------------------|----------------------------------|------------------------------|---------------------------------------|
| Logistic regression | Churn/response prediction        | Interpretable probabilities  | Limited structure nonlinear           |
| Decision tree       | Retention and targeting rules    | Easy to explain              | Can overfit and be unstable           |
| Random forest       | Churn/propensity scoring         | Handles interactions; robust | Lower transparency                    |
| SVM                 | High-dimensional classification  | Strong nonlinear boundary    | Tuning and explanation difficulty     |
| KNN                 | Similarity-based prediction      | Simple local learning        | Scaling sensitive; slow at prediction |
| K-means             | Customer segmentation            | Simple behavioral grouping   | Needs k; sensitive to scale           |
| Association rules   | Cross-sell and affinity analysis | Readable IF-THEN patterns    | Rule explosion; not probabilistic     |

## Methodology

### 1. Research Design and Data Structure

The proposed design treats CRM analysis as a pipeline rather than a single modeling step. A customer-level analytical file is created by joining profile, transaction, campaign, service, and complaint records at a defined observation date. Core behavioral variables include recency, purchase frequency, monetary value, relationship tenure, complaint count, satisfaction score, discount dependence, service contacts, and support-call frequency. Categorical fields such as channel, product category, and region can be one-hot encoded. Personally, identifying fields are excluded because they do not improve the methodological comparison and create unnecessary privacy risk.

For a predictive retention experiment, the target is binary: 1 represents a customer who churns within the defined prediction window and 0 represents a retained customer. Records are split using stratification so the churn proportion is similar in training and test sets. Continuous variables are standardized for distance- and margin-based models such as KNN and SVM, while tree models can use the unscaled values. Missing values should be imputed using rules fitted only on the training data to prevent leakage. Feature selection is performed before final model fitting, reflecting the importance of preparation emphasized by Moro et al. (2014) and Coussemont et al. (2017).

### 2. Techniques Compared

Five supervised models are used for the numerical benchmark: logistic regression, decision tree, random forest, k-nearest neighbors, and support vector machine. Logistic regression provides a transparent baseline and calibrated propensity scores. Decision trees generate manager-readable rules. Random forests reduce the variance of individual trees and capture interactions. KNN tests similarity-based classification, while SVM represents a margin-based nonlinear classifier. Neural networks are discussed from the literature but are not included in the compact demonstration because their performance depends strongly on architecture and tuning, which would expand the study beyond the intended six-to-seven-page scope.

Two unsupervised/descriptive techniques are evaluated conceptually alongside the predictive models. K-means clustering is applied to standardized recency-frequency-monetary and service variables to identify behavioral segments such as high-value loyal, promotion-sensitive, low-engagement, and at-risk customers. Association-rule mining is applied to discretized transaction and product indicators to identify co-occurrence patterns useful for cross-selling or campaign design. These techniques should not be ranked using churn accuracy because they solve different CRM problems; instead, their outputs are assessed for cluster cohesion, stability, support, confidence, lift, and managerial interpretability.

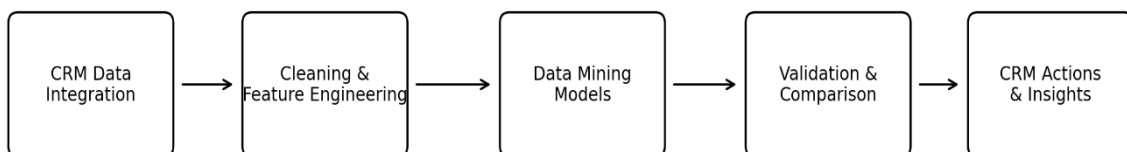


**Figure 1.** Proposed comparative data-mining workflow for CRM

### 3. Benchmark Procedure and Evaluation Metrics

Because no proprietary organization dataset was supplied, the numerical results in this methodology paper are produced from a fixed-seed synthetic CRM benchmark of 5,000 customers. The simulation reproduces common CRM relationships: churn probability increases with long recency, complaints, support calls, low satisfaction, and heavy discount dependence, and decreases with longer tenure, higher purchasing frequency, and greater monetary value. The outcome contains approximately one-quarter churners, preserving a realistic class-imbalance challenge. The dataset is divided into 70% training and 30% testing records using a fixed random seed of 42. The same split is used for every model so the comparison is controlled and reproducible.

Model performance is measured using accuracy, precision, recall, F1-score, and AUC. Accuracy measures overall correctness but can favor the majority class. Precision measures the proportion of predicted churners that actually churn, while recall measures the proportion of actual churners detected. F1-score balances

precision and recall. AUC measures ranking quality across thresholds and is useful when a campaign can contact only a fraction of customers. In a deployed CRM system, lift, expected profit, campaign capacity, and retention cost should also be reported, consistent with the business-oriented perspective of Neslin et al. (2006) and Burez and Van den Poel (2009).

### Results

Table 2 reports the controlled benchmark. The values are simulation outputs, not claims about a specific company. KNN obtains the highest raw accuracy (0.764) but detects only 14.8% of churners, illustrating why accuracy alone is unsuitable for an imbalanced retention problem. Logistic regression obtains the highest AUC (0.711) and recall of 0.635, making it the strongest ranking model in this particular simulation. Random forest produces the second-highest accuracy and a more balanced precision-recall profile than KNN. The decision tree is easy to explain but has the weakest overall discrimination, while the SVM performs between the tree and logistic alternatives.

**Table 2.** Predictive benchmark on the synthetic CRM dataset

| Technique              | Accuracy | Precision | Recall | F1    | AUC   |
|------------------------|----------|-----------|--------|-------|-------|
| Logistic Regression    | 0.675    | 0.407     | 0.635  | 0.496 | 0.711 |
| Decision Tree          | 0.593    | 0.332     | 0.608  | 0.430 | 0.647 |
| Random Forest          | 0.727    | 0.462     | 0.508  | 0.484 | 0.695 |
| K-Nearest Neighbors    | 0.764    | 0.636     | 0.148  | 0.240 | 0.645 |
| Support Vector Machine | 0.665    | 0.390     | 0.587  | 0.469 | 0.685 |

The comparative interpretation agrees with the pre-2018 literature in two respects. First, there is no universal best technique across datasets and performance measures; model choice changes when the objective shifts from maximum accuracy to detecting more at-risk customers (Keramati et al., 2014; Vafeiadis et al., 2015). Second, a relatively simple model can remain competitive when features and preprocessing are carefully designed (Coussement et al., 2017). For CRM deployment, the recommended procedure is therefore to select a model at the operating threshold that maximizes the relevant business objective, then combine its scores with customer value and segment information. A high-risk, high-value customer may justify an expensive retention action, whereas a low-value customer with the same risk score may receive an automated offer.

### Conclusion and Discussion

This paper presented a methodology for comparing data mining techniques for customer relationship management using both predictive and descriptive perspectives. The literature up to 2017 shows that classification, clustering, association analysis, and behavioral feature engineering have all contributed to CRM applications, but they serve different managerial purposes. The controlled benchmark further demonstrates that overall accuracy may be deceptive when the target class is uncommon. In the reported simulation, logistic regression provided the best AUC and churn recall, while KNN achieved higher accuracy primarily because it favored the majority class. These results should be interpreted as a demonstration of the evaluation procedure rather than as evidence that one algorithm will dominate on real organizational data. For a full empirical study, the same pipeline should be applied to an organization-specific dataset or a clearly

documented public CRM dataset, with temporal validation where possible. Future work can add cost-sensitive learning, gradient boosting, neural networks, uplift modeling, survival analysis, and profit-based evaluation. The central methodological conclusion is that technique selection should jointly consider predictive discrimination, minority-class detection, interpretability, computational cost, data preparation, and the CRM action that follows the prediction. Such a balanced framework converts data mining from an isolated technical exercise into a decision-support component of customer relationship management.

## References

- Bose, I., & Mahapatra, R. K. (2001). Business data mining—A machine learning perspective. *Information & Management*, 39(3), 211–225. DOI: 10.1016/S0378-7206(01)00091-X.
- Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, 164(1), 252–268. DOI: 10.1016/j.ejor.2003.12.010.
- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626–4636. DOI: 10.1016/j.eswa.2008.05.027.
- Chen, M.-C., Chiu, A.-L., & Chang, H.-H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4), 773–781. DOI: 10.1016/j.eswa.2004.12.033.
- Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95, 27–36. DOI: 10.1016/j.dss.2016.11.007.
- Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, 26(2), 181–188. DOI: 10.1016/S0957-4174(03)00133-7.
- Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24, 994–1012. DOI: 10.1016/j.asoc.2014.08.041.
- Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, 19(1/2), 65–93. DOI: 10.1108/JSIT-10-2016-0061.
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22–31. DOI: 10.1016/j.dss.2014.03.001.
- Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering*, 109, 233–252. DOI: 10.1016/j.cie.2017.04.017.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43(2), 204–211. DOI: 10.1509/jmkr.43.2.204.
- Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), 2592–2602. DOI: 10.1016/j.eswa.2008.02.021.
- Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, 24(4), 483–502. DOI: 10.1016/S0160-791X(02)00038-6.
- Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, 31(1), 127–137. DOI: 10.1016/S0167-9236(00)00123-8.
- Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 1–9. DOI: 10.1016/j.simpat.2015.03.003.
- Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications*, 23(2), 103–112. DOI: 10.1016/S0957-4174(02)00030-1