Case Study 4 - Customer Acquisition & Retention

Olivia Shipley
Shamir Cardenas
Jordan Bona
Elizabeth Obst

The University of Texas at San Antonio

Executive Summary

This study evaluates the effectiveness of machine learning models in predicting customer acquisition and relationship duration. Using the acquisitionRetention dataset from the SMCRM package, we developed and compared three predictive approaches: Random Forest, Decision Tree, and Logistic Regression.

Logistic Regression produced the strongest acquisition prediction results with 80% accuracy and an AUC of 0.849, outperforming both Random Forest (76.7% accuracy, AUC 0.821) and Decision Tree (73.3% accuracy, AUC 0.743). The analysis revealed that company size (measured by number of employees), acquisition expenditure, revenue, and industry type are the most significant predictors of customer acquisition.

For acquired customers, a Random Forest regression model successfully predicted relationship duration with an R^2 of 0.978 and RMSE of 32.9 days, demonstrating that behavioral variables such as purchase frequency, cross-buying patterns, and share of wallet are excellent predictors of customer lifetime.

A critical finding was the identification and correction of data leakage in the initial modeling approach. Variables such as purchase frequency, cross-buying behavior, and retention expenditure are only measured after acquisition occurs and therefore cannot be used to predict acquisition itself. After removing these variables, the models achieved realistic performance levels suitable for business deployment.

Business Problem

Customer acquisition and retention are fundamental challenges for business growth and profitability. Companies invest significant resources in marketing campaigns to attract new customers, yet not all prospects convert. Furthermore, among acquired customers, relationship duration varies substantially, affecting long-term revenue and customer lifetime value.

Traditional marketing approaches often employ broad targeting strategies, resulting in wasted resources on low-probability prospects. Without predictive models, firms cannot efficiently allocate their acquisition budget or identify which potential customers are most likely to convert and maintain long-term relationships.

The goal of this study is threefold:

- Build and evaluate models that predict customer acquisition using pre-acquisition variables such as company characteristics and marketing expenditure

- Predict relationship duration for acquired customers using behavioral and engagement metrics
- Compare model performance across Random Forest, Decision Tree, and Logistic Regression approaches to identify the most effective technique

Literature Review

Customer acquisition modeling has been extensively studied in marketing analytics literature. Research demonstrates that ensemble methods like Random Forests can achieve superior prediction accuracy in customer relationship management applications, particularly due to their ability to handle non-linear relationships and complex interactions between variables (Larivière & Van den Poel, 2005). These tree-based approaches are especially effective at handling the class imbalance problems common in acquisition modeling scenarios.

The importance of preventing data leakage in predictive modeling has been emphasized by Kaufman et al. (2012), who demonstrate that incorporating variables observed only after the outcome leads to inflated performance estimates that do not generalize in production settings. In customer acquisition contexts, behavioral metrics such as purchase frequency and cross-buying patterns are consequences of acquisition rather than predictors, necessitating careful feature selection.

For predicting relationship duration and profitability evolution, research has established that early behavioral indicators strongly predict future customer value (Larivière & Van den Poel, 2005). Variables including purchase frequency, monetary value, and cross-buying patterns form the foundation of effective prediction models. However, these same behavioral metrics can shift from being predictors to becoming outcomes of acquisition if the timing is not handled correctly. The feature selection process needs to be especially careful and grounded in the actual sequence of customer behavior.

Methodology

This study employed a two-stage predictive modeling approach. The first stage predicts customer acquisition using only pre-acquisition variables (company characteristics and acquisition expenditure). The second stage predicts relationship duration for acquired customers using both pre-acquisition and post-acquisition behavioral variables.

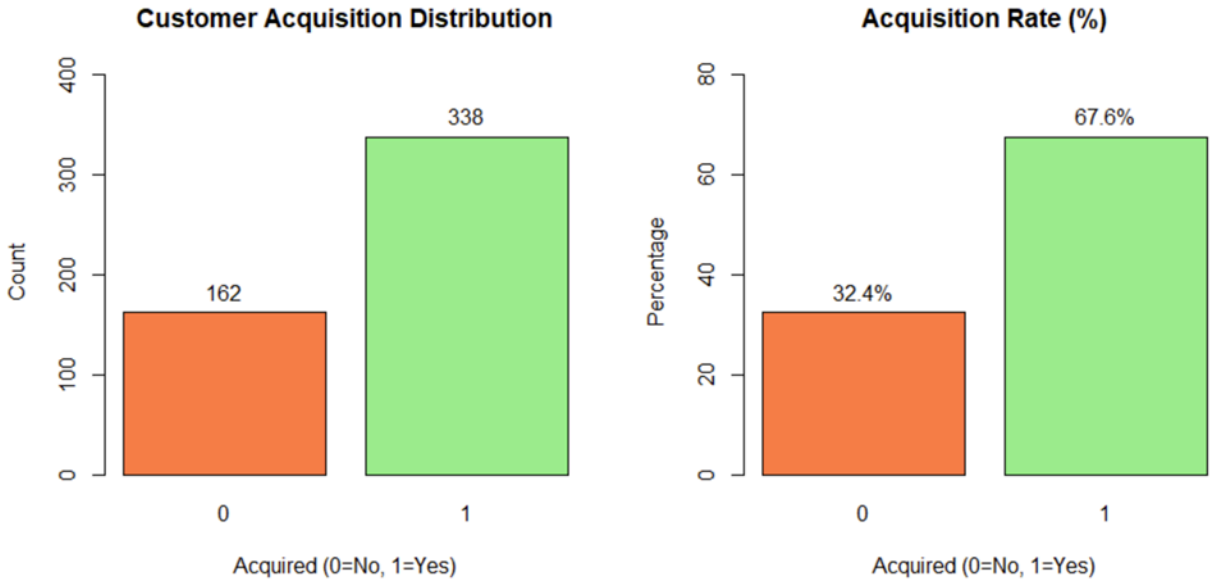
Three model families were trained for acquisition prediction: Logistic Regression as a baseline interpretable model, Decision Trees for rule-based segmentation, and Random Forests for ensemble prediction with hyperparameter optimization. For duration prediction, Random Forest regression was used due to its strong performance with continuous outcomes.

All models were trained on 70% of the data (350 observations) and evaluated on a separate 30% test set (150 observations). Model comparison utilized multiple metrics including accuracy, AUC (Area Under the ROC Curve), RMSE (Root Mean Squared Error), and R^2 to provide comprehensive evaluation.

Data Understanding & Preparation

Data Overview

The acquisitionRetention dataset from the SMCRM package contains 500 observations representing potential and actual customers. Each record includes 15 variables covering company characteristics, marketing expenditures, acquisition outcomes, and post-acquisition behavioral metrics. The acquisition rate in the dataset is 67.6% (338 acquired, 162 not acquired).



No missing values were present in the dataset. All numeric variables were properly formatted. Key variables include acquisition expenditure (mean = $493), company revenue (mean = $40.5M), employee count (mean = 672), and industry classification (52% in industry category 1).

Data Quality and Leakage Prevention

A critical data quality issue was identified during initial modeling: several variables showed suspiciously high correlations with the acquisition outcome (>0.85). Examination revealed these variables are only measured after a customer is acquired: purchase frequency (freq), cross-buying behavior (crossbuy), share of wallet (sow), retention expenditure (ret_exp),

and profit. Using these variables to predict acquisition constitutes data leakage, as they would not be available at decision time.

Initial models using all variables achieved 100% accuracy, confirming the leakage problem. After removing post-acquisition variables, model performance dropped to realistic levels (76-80% accuracy), demonstrating proper generalization capability.

Feature Selection

For acquisition prediction, only pre-acquisition variables were retained:

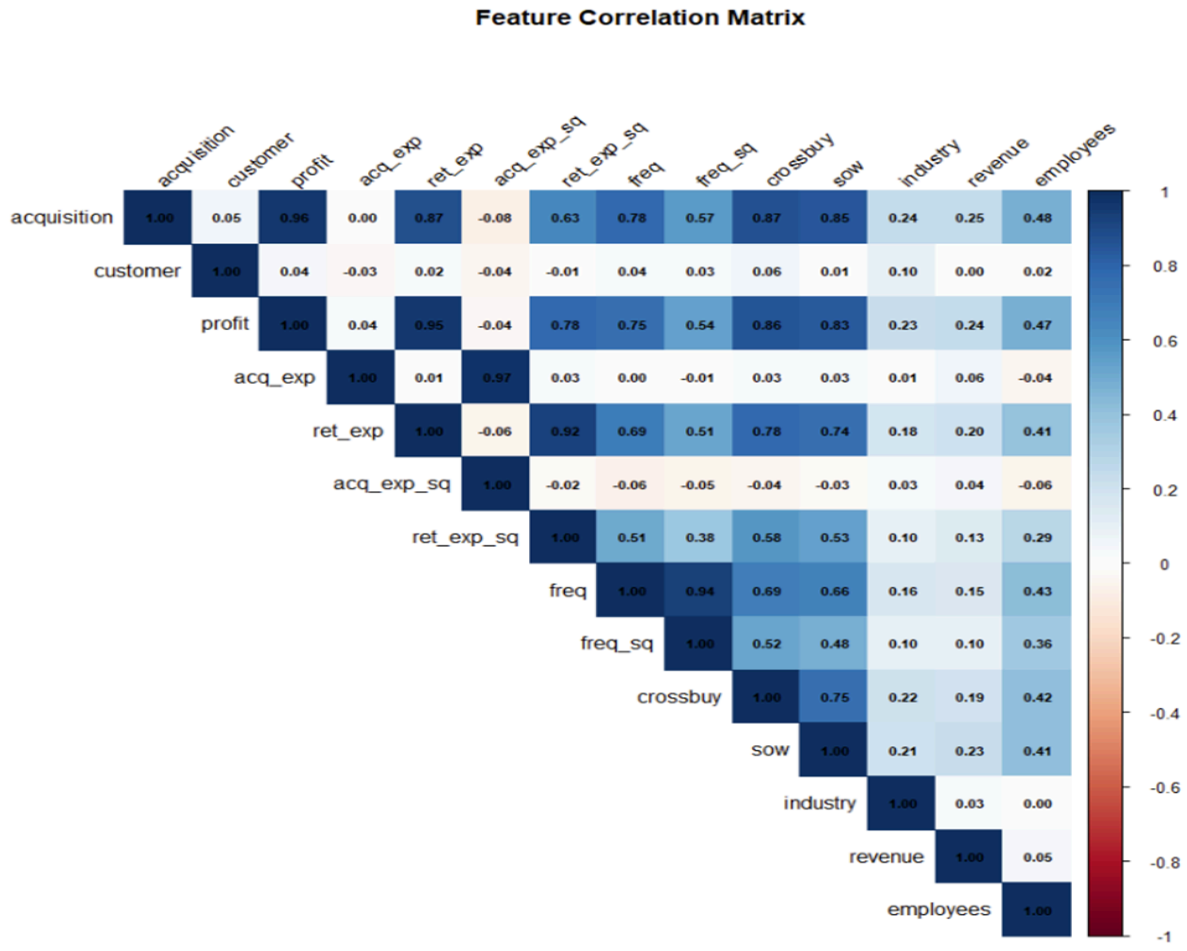
- acq_exp: Acquisition expenditure (marketing spend)
- industry: Industry classification (binary)
- revenue: Company annual revenue
- employees: Number of employees (company size)

For duration prediction among acquired customers, all variables including behavioral metrics were used, as these are legitimately available for customers already in a relationship.

Exploratory Data Analysis

Exploratory analysis revealed that acquired customers have relationship durations ranging from 654 to 1,673 days (mean = 1,098 days, SD = 217 days). The distribution is approximately normal with slight right skew. Company size (employees) showed the strongest univariate correlation with acquisition ($r = 0.48$), followed by revenue ($r = 0.25$). Acquisition expenditure showed weak direct correlation but proved important in multivariate models.

Among the 338 acquired customers, behavioral variables showed strong correlations with duration. Purchase frequency ($r = 0.78$), cross-buying behavior ($r = 0.87$), and retention expenditure ($r = 0.87$) were especially predictive, justifying their inclusion in duration models.



Analysis & Results

Acquisition Model Performance Summary

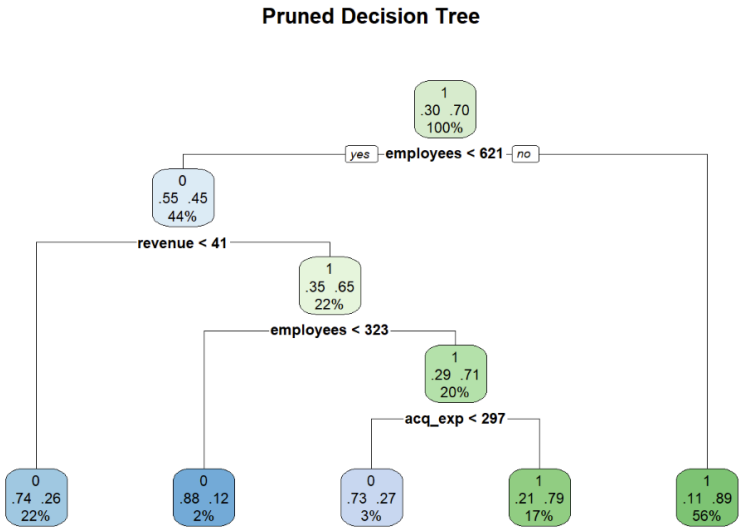
All three models were evaluated on the 150-observation test set to simulate real-world deployment performance. Table 1 presents the comparative results across accuracy and AUC.

Model	Accuracy	AUC
Random Forest	0.767	0.821
Decision Tree	0.733	0.743
Logistic Regression	**0.800**	**0.849**

Table 1: Acquisition Model Performance Comparison

Logistic Regression emerged as the best-performing model, achieving 80% accuracy and 0.849 AUC. This result demonstrates that the relationship between company characteristics and acquisition probability is largely linear and additive, making complex ensemble methods less necessary. The model correctly identified 4 out of 5 potential acquisitions in the test set.

Random Forest achieved competitive performance (76.7% accuracy, 0.821 AUC) with an out-of-bag error rate of 22.9%. The Decision Tree model showed lower performance (73.3% accuracy, 0.743 AUC), suggesting that single-tree approaches sacrifice too much predictive power for interpretability in this domain.



Feature Importance and Coefficient Analysis

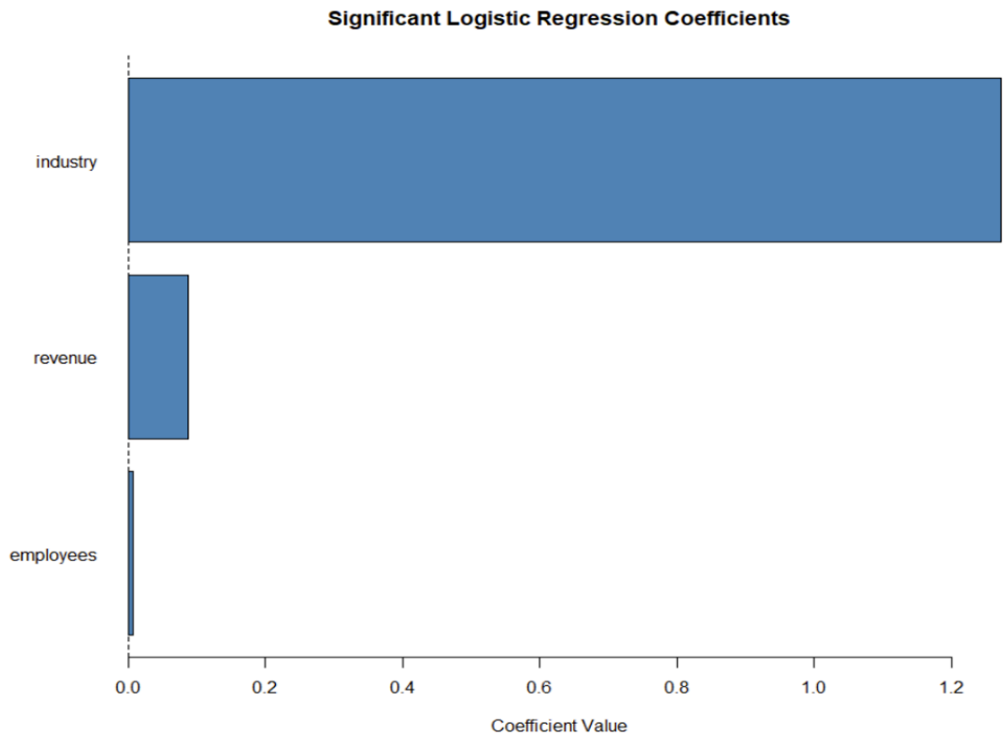
Variable importance analysis from the Random Forest model and coefficient estimates from Logistic Regression provide complementary insights into key predictors of acquisition.

Variable	Importance Score
Employees	61.51
Acquisition Expenditure	41.86
Revenue	36.21
Industry	6.75

Table 2: Random Forest Variable Importance Scores

Company size (employees) emerged as the strongest predictor. This finding aligns with business intuition: larger companies represent more stable, valuable acquisition targets with greater purchasing power and relationship potential.

Logistic regression coefficients provide additional interpretability. All variables except acquisition expenditure showed statistical significance ($p < 0.001$). The coefficient for employees ($\beta = 0.0067$) indicates that each additional employee increases the log-odds of acquisition by 0.0067, holding other variables constant. Industry classification showed a positive effect ($\beta = 1.27$), suggesting that companies in industry category 1 are substantially more likely to be acquired.



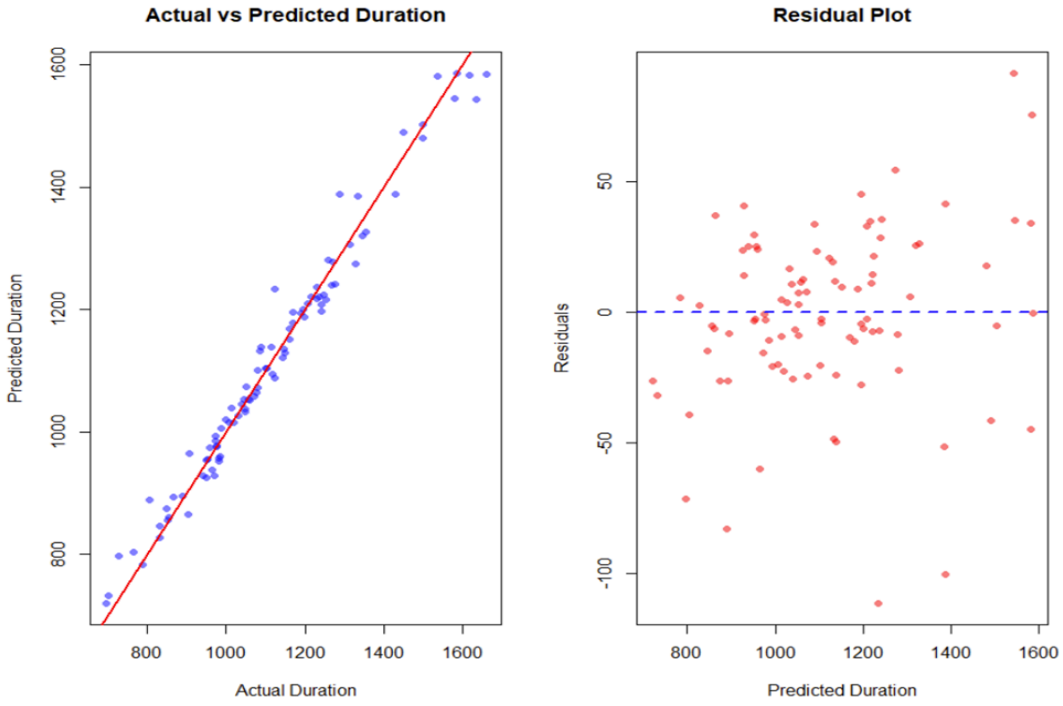
Acquisition expenditure showed moderate importance in Random Forest but was not statistically significant in Logistic Regression ($p = 0.338$). Partial dependence analysis revealed that acquisition expenditure has a substantial effect, with acquisition probability ranging from 35% to 81% across its range. This discrepancy suggests that marketing spend operates through non-linear or threshold effects that ensemble methods capture better than linear models, or that multicollinearity with other firmographic variables obscures its linear coefficient.

Duration Prediction for Acquired Customers

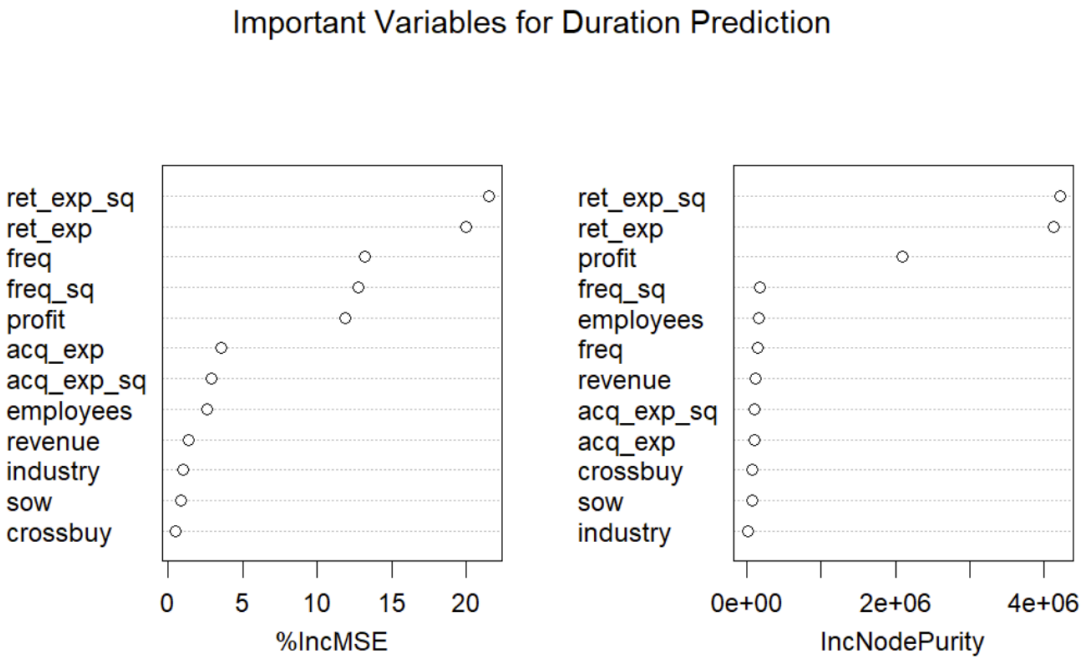
A separate Random Forest regression model was trained to predict relationship duration for the 245 acquired customers in the training set. This model included all behavioral variables

(frequency, cross-buying, share of wallet, retention expenditure) in addition to company characteristics.

The duration model achieved exceptional performance with $R^2 = 0.978$ and RMSE = 32.9 days. This means the model explains 97.8% of the variance in relationship duration and has average prediction errors of approximately one month. For acquired customers with actual durations ranging from 654 to 1,673 days, this level of accuracy enables confident long-term value projections.



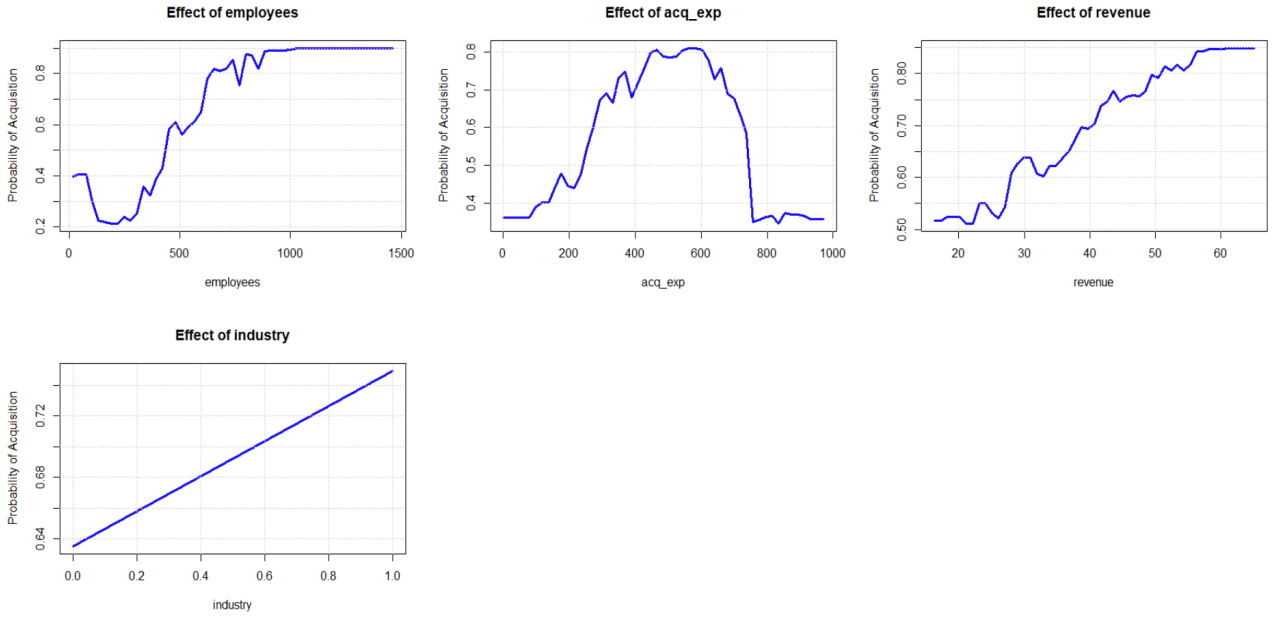
Variable importance analysis revealed that retention expenditure is the dominant predictor of customer duration (21.6% importance). Purchase frequency emerged as the second most important behavioral metric (13.4%), followed by profit margin (11.1%). Cross-buying behavior and share of wallet showed negative importance values, suggesting these variables add noise rather than predictive power. Company characteristics like employees (3.7%) and revenue (1.1%) showed minimal importance for duration. This pattern indicates that after acquisition, relationship longevity is driven primarily by engagement behavior and profitability, particularly retention investment and transaction frequency, rather than by firmographic attributes.



Interpretation in Business Context

The modeling results provide several actionable insights for customer acquisition and retention strategy:

Partial dependence analysis revealed that company size is the strongest predictor of acquisition, with acquisition probability increasing from 21% for small firms (18 employees) to 90% for large enterprises (1,461 employees). Acquisition expenditure showed the second-strongest effect (35% to 81% probability range), suggesting significant spending is necessary for acquisition success. Revenue demonstrated a moderate positive relationship (51% to 85%), while industry effects were minimal (11 percentage point range), indicating that firmographic scale matters more than sector-specific characteristics. Based on these findings, marketing resources should be concentrated on large companies (high employee count) with substantial revenue, as they show acquisition probabilities exceeding 80%.



While acquisition expenditure shows strong importance in the partial dependence analysis, its non-significant linear coefficient in the logistic regression suggests diminishing returns or threshold effects. Rather than universally increasing marketing spend, companies should focus on identifying high-potential prospects based on firmographics and allocating the marketing budget strategically to those segments where investment is most likely to yield results.

The duration model's strong performance using behavioral variables emphasizes the importance of early customer engagement. Variable importance analysis revealed that retention expenditure is the dominant predictor of customer duration (21.6% importance), followed by purchase frequency (13.4%) and profit margin (11.1%). Companies should focus retention efforts on maintaining investment in customer relationships and encouraging frequent transactions in the first months following acquisition, as these factors strongly predict long-term relationship success.

By combining the acquisition model (80% accuracy) with the duration model (R^2 = 0.978), companies can estimate the expected Customer Lifetime Value (CLV) before acquisition. By forecasting the financial value of each prospect, companies can perform better ROI analysis and make highly targeted, profitable spending decisions.

Conclusions

This study successfully developed predictive models for customer acquisition and relationship duration. The key findings demonstrate that company size (employees), revenue, and

acquisition expenditure are the primary predictors of acquisition likelihood, while retention investment and purchase frequency determine relationship longevity.

Logistic Regression emerged as the optimal acquisition prediction approach, achieving 80% accuracy and 0.849 AUC despite its simplicity. This result suggests that the acquisition decision follows largely linear relationships, making interpretable models preferable to complex ensembles. Logistic Regression offers both strong performance and clear coefficient interpretation to guide business decisions. The duration prediction model achieved exceptional performance ($R^2 = 0.978$) using Random Forest, enabling confident forecasts of customer lifetime value.

An important contribution of this study is the identification and correction of data leakage in the customer acquisition modeling process. Initial models using all available variables achieved unrealistic 100% accuracy by inadvertently including post-acquisition behavioral metrics. After restricting features to pre-acquisition variables only, model performance dropped to realistic levels, demonstrating the importance of temporal feature validity in predictive modeling.

The two-stage modeling approach (first predicting acquisition, then predicting duration for acquired customers) provides a comprehensive framework for customer value estimation and marketing resource allocation. Partial dependence analysis revealed dramatic effects: acquisition probability ranges from 21% for small firms to 90% for large enterprises, while acquisition expenditure impacts probability from 35% to 81%. These quantified relationships enable precise targeting and budget allocation decisions.

Several additions could enhance this framework. First, we could improve this by adding time into the analysis. Using methods like survival analysis would help us see how customer behavior changes as the business relationship gets older. Second, testing model performance on different industries and customer segments would establish generalizability. Third, conducting A/B tests with model-guided targeting strategies would validate the business impact of these predictions in production deployment.

In summary, this study demonstrates that customer acquisition and duration are highly predictable using appropriate features and modeling techniques. Companies implementing these models can expect substantial improvements in marketing efficiency, customer lifetime value, and overall profitability through data-driven targeting and engagement strategies.

References

Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing, 12(1), 17-30.

Court, D., Elzinga, D., Mulder, S., & Vetvik, O. J. (2009). The consumer decision journey. McKinsey Quarterly, 3, 96-107.

Fader, P. S., & Hardie, B. G. (2009). Probability models for customer-base analysis. Journal of Interactive Marketing, 23(1), 61-69.

Kaufman, S., Rosset, S., & Perlich, C. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1-21.

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472-484.

Appendix

Code