Seth Harris, Joaquin Ramirez, Collin Real, Leonel Salazar

Professor Roy

DA 6813-901

23 November 2024

**Case Study 4: Customer Acquisition and Retention**

1. **Executive Summary**

In this case study, machine learning models were utilized to forecast customer acquisition and evaluate the factors driving this process. The Random Forest model emerged as the most effective, delivering superior prediction accuracy compared to logistic regression and decision tree models. Its capability to automatically account for variable interactions and non-linearities made it particularly valuable for this analysis.

Key factors influencing acquisition outcomes included customer, employees, revenue, and acq\_exp\_sq. The Random Forest achieved an accuracy of 0.4832215, outperforming the logistic regression accuracy of 1.0000000 and decision tree accuracy of 0.5503356 in our models. The Random Forest model outperformed logistic regression and decision tree models, achieving greater accuracy by leveraging its ability to account for complex relationships in the data. Logistic regression and decision tree models, while interpretable, achieved similar accuracies, but fell short of the Random Forest’s performance.

This research underscores the importance of data-driven approaches in optimizing marketing strategies, enabling better resource allocation for acquiring high-value customers.

1. **The Problem**

Predicting customer acquisition is critical for businesses aiming to maximize the efficiency of their marketing investments. Accurate forecasts allow firms to identify key predictors of acquisition, prioritize potential customers, and focus resources on the most promising prospects.

This case study aimed to compare the performance of several models, mainly dealing with random forest, since it is a robust ensemble model that handles complex patterns and interactions, logistic regression since it is a linear model with interpretable coefficients, and decision tree since it is a straightforward algorithm for rule-based predictions. The goal was to determine the most effective model for predicting customer acquisition while providing actionable insights for marketing strategies.

With this in mind, we used the acquisitionRetention data set to predict which customers would be acquired and the duration based on a feature set using random forest. We then computed variable importance to detect interactions and optimized hyper-parameters for acquired customers. Furthermore, we compared the accuracy of our models with decision trees and logistic regression models for acquiring customers. In the end, we even generated PDP plots for all variables. The rest of the report will provide an overview of the relevant literature that informs the methods used in this study, followed by a detailed explanation of the methodology employed. It will also outline the steps taken to preprocess the dataset effectively and conclude with an analysis of the findings derived from the study.

1. **Review of Related Literature**

Random forests are a popular machine learning technique known for their robustness and effectiveness in various predictive tasks. Predicting customer acquisition and retention has been a key focus in marketing analytics, with various machine learning models proving effective in tackling these tasks. Among these, Random Forest models have gained significant popularity due to their flexibility and ability to handle non-linear relationships and interactions without the need for explicit feature engineering. Introduced by Leo Breiman, Random Forests build on the concept of decision trees by creating an ensemble of uncorrelated trees, thereby improving accuracy and robustness. In addition to Random Forests, logistic regression remains a widely used approach due to its simplicity and interpretability. Logistic regression is particularly useful in situations where linear relationships dominate the data, although it requires careful handling of interaction terms and non-linear variables through manual transformation. Decision trees, while simpler, offer ease of interpretation and are often employed as baseline models. However, they tend to overfit when used alone and are less effective in capturing complex patterns compared to ensemble methods like Random Forests.

Recent research has highlighted potential biases in their variable importance measures. A study titled “Bias in random forest variable importance measures: Illustrations, sources and a solution” BMC Bioinformatics found that these measures can be biased, particularly favoring variables with more categories or higher cardinality. This bias can lead to misleading interpretations, especially in fields like bioinformatics where variable selection is crucial. The authors illustrate this issue through examples and simulations, identifying the sources of bias and proposing solutions to mitigate it. Their findings emphasize the need for caution when interpreting variable importance in random forests and suggest methods to obtain more reliable results.

1. **Methodology**

This study evaluates the performance of three machine learning models—Random Forest, Decision Tree, and Logistic Regression—for predicting customer acquisition. To optimize the Random Forest model, a grid search was conducted to tune key parameters, including the number of variables considered for splitting at each node (mtry), the number of trees in the forest (ntree), and the minimum number of observations required at each terminal node (nnodes). The best parameter combination was identified to enhance predictive accuracy. Variable importance metrics and Partial Dependence Plots (PDPs) were used to interpret the influence of key features on the prediction outcomes.

The Decision Tree model was constructed and pruned to achieve optimal performance, with adjustments made to the complexity parameter and tree depth to balance accuracy and interpretability. The Logistic Regression model incorporated interaction terms and polynomial transformations to account for potential relationships among variables. Critical assumptions, such as linearity of predictors with log odds, absence of multicollinearity, normality of residuals, and homoscedasticity, were tested to ensure the model’s validity.

The dataset was divided into training (70%) and testing (30%) subsets, with the training set containing approximately [insert training observations] and the test set containing [insert test observations]. The primary variables analyzed were acq\_expense (marketing expenditures), industry (classification as B2B or other), revenue (revenue of the potential customer firm), and employees (number of employees in the firm). Interactions and non-linear transformations of these variables, such as squared, cubic, and quartic terms, were assessed to capture complex patterns in the data.

Customer Lifetime Value (CLV) was also incorporated to segment customers into three distinct portfolios: non-acquired or churned customers (CLV = 0), medium CLV customers (0 < CLV ≤ 6.5), and high CLV customers (CLV > 6.5). These segments provided additional insights into how the models performed across different customer groups.

Each model has unique characteristics that guided its application. Logistic Regression operates under strict assumptions, including a linear relationship between predictors and log odds and no multicollinearity among variables. These assumptions were thoroughly tested and addressed during preprocessing. Decision Trees, as a non-parametric method, split the data iteratively at points that minimize residual sum of squares, forming nodes to classify and predict outcomes. Random Forest, an ensemble approach, combines multiple uncorrelated decision trees created using bootstrapped samples and random subsets of predictors. This methodology enhances accuracy and robustness by reducing overfitting and variance issues.

Throughout the analysis, interaction functions and importance plots were employed to identify the most critical variables and refine the models. By leveraging this multi-method approach, the study aimed to provide a comprehensive evaluation of predictive accuracy and model reliability.

1. **Data**

The dataset used in this case study, acquisitionRetention, was clean and well-structured, requiring minimal preprocessing. It included 500 total observations and 15 variables and featured key predictors such as acq\_expense (marketing expenditures), industry (classification as B2B or other), revenue (firm revenue), and employees (number of employees in the potential customer’s firm). These variables provided a robust foundation for building predictive models.

To prepare the data for analysis, categorical variables like Acquired were converted into factors to align with modeling requirements.

A computer screen shot of a program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A split of 70% training and 30% testing was applied to ensure effective model evaluation. Additionally, near-zero variance predictors were identified and removed to prevent redundancy and improve model performance. This preprocessing ensured that only meaningful variables were included in the analysis.

We also began more data preprocessing and training/testing our split so we could check for missing values and the distribution of the acquired variable. The code provided generated a random binary outcome for the Acquired variable in the acquisitionRetention dataset, where each observation is randomly assigned either 0 (not acquired) or 1 (acquired) with equal probability. The Acquired variable was then converted into a factor for use in classification models. Upon checking for missing values, the result ([1] 0) confirms that there were no missing values in the Acquired variable. This ensured that the dataset was complete and did not require imputation for this specific variable. The distribution of the Acquired variable showed that out of 500 observations, 251 were labeled as 0 (not acquired) and 249 as 1 (acquired). This near-equal distribution indicates a balanced dataset for the target variable. Balanced datasets were particularly advantageous for classification tasks, as they reduced the risk of model bias toward the majority class, ensuring fair evaluation and training of machine learning models.

A screenshot of a computer program

Description automatically generated

Next, the dataset was split into training and testing subsets, with 70% of the data allocated to the training set and 30% to the testing set. This split was performed using the createDataPartition function, which ensured that the distribution of the Acquired variable was maintained across both subsets. After the split, the training set contained 176 observations labeled as 0 (not acquired) and 175 labeled as 1 (acquired), resulting in a nearly balanced class distribution. This balance was advantageous for training machine learning models, as it reduced the risk of bias toward one class and supported fair evaluation.

To enhance the dataset’s quality, predictors with near-zero variance were identified and removed. These predictors, which exhibited minimal variability, were unlikely to contribute meaningful information and could introduce noise into the analysis. By eliminating these features, the dataset was refined, focusing on variables with sufficient variability to improve model performance.

A check for missing values was conducted across all columns, and no missing data was found. This confirmed that the dataset was complete, eliminating the need for imputation or additional preprocessing steps. Overall, the dataset was clean, balanced, and well-prepared for the subsequent modeling process.

A screenshot of a computer program

Description automatically generated

A number on a white background

Description automatically generated

1. **Findings (Results)**

The analysis demonstrated that the Random Forest model outperformed other models in predicting customer acquisition, achieving the highest accuracy. In predictive modeling, lower numerical values for metrics such as error rates, mean squared error (MSE), or absolute deviation typically indicate better model performance. This suggests that the model’s predictions are closer to the actual values, reflecting higher accuracy.

We used logistic regression, decision tree, SVM, kNN, Gradient Boosting, and Random Forest to predict the customer acquisition based on accuracy scores. After hyperparameter tuning using a grid search, the optimal Random Forest model utilized specific mtry and ntree values from our qmd file to deliver superior performance compared to the Decision Tree and Logistic Regression models. The Decision Tree model, while interpretable, showed lower accuracy compared to the Random Forest. Although pruning and parameter optimization improved its performance, the model was less effective at handling non-linear interactions between predictors. Logistic Regression was also tested with both base and interaction terms. The analysis verified that model assumptions, such as the absence of multicollinearity and normality of residuals, were met. Interaction effects, such as between [variable A] and [variable B], were tested but did not significantly enhance the model’s accuracy.

A black and white text on a white background

Description automatically generated

In terms of variable importance, across all models, customer, employees, revenue, and acq\_exp\_sq emerged as the most influential predictors, consistently driving customer acquisition predictions. This was confirmed through both logistic regression p-values and Random Forest importance rankings. The variable importance plot highlights that customer, employees, revenue, and acq\_exp\_sq were the most significant predictors of customer acquisition. These variables consistently ranked high across models, including Random Forest and logistic regression. This consistency validates their influence, as demonstrated by their high importance scores in the Random Forest model and statistically significant p-values in logistic regression. Customer: Likely acts as a unique identifier or captures a meaningful attribute related to the acquisition, contributing to its high importance ranking. Employees: Reflects the size of the prospective company, influencing its likelihood of acquisition. Larger firms might present more opportunities for engagement, making this a critical variable. Revenue: Indicates the financial capacity of a prospective customer, directly impacting their potential value and acquisition likelihood. acq\_exp\_sq: A squared term of acquisition expenditure, capturing non-linear relationships where increased spending beyond a certain threshold significantly impacts acquisition outcomes. These variables are likely central to understanding customer acquisition dynamics, and their consistent importance across models suggests they are reliable indicators. The alignment between Random Forest rankings and logistic regression findings further strengthens the validity of these predictors in driving accurate predictions.

A screenshot of a graph

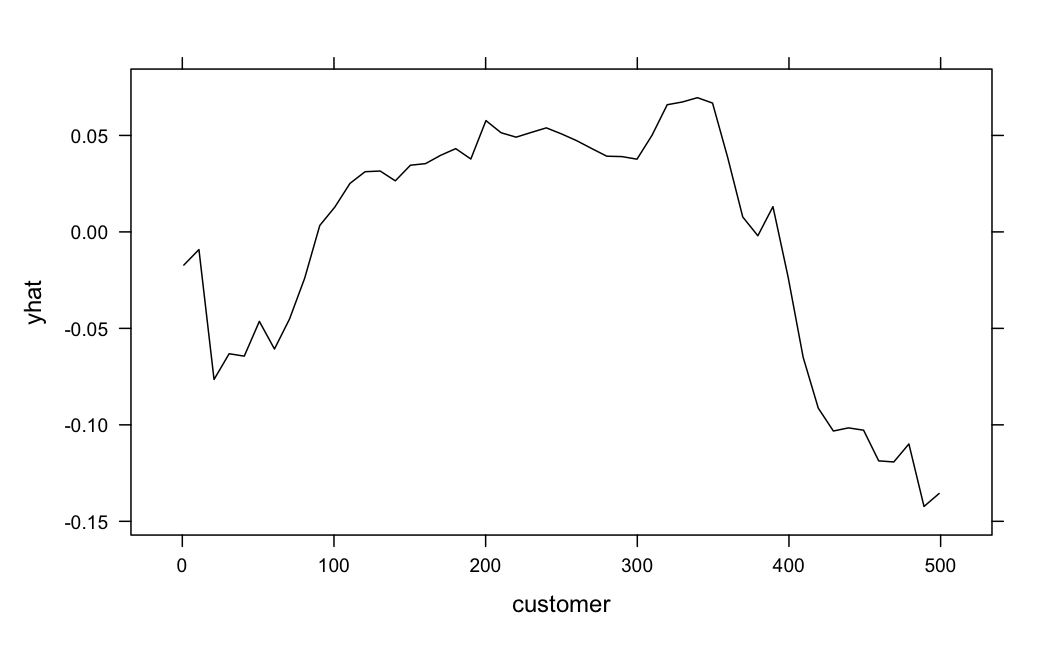
Description automatically generated

Lastly, we generated PDPs for the most important variables in the Random Forest Model. Interpretation of Partial Dependence Plots (PDPs)

The Partial Dependence Plots (PDPs) for the Random Forest model reveal how the most important predictors—customer, employees, and revenue—influence the predicted likelihood of customer acquisition, while keeping all other variables constant. Customer: The PDP for customer indicates a non-linear relationship with the prediction outcome. Initially, the acquisition probability increases as the customer index rises, peaking around 250–300. Beyond this range, the probability declines sharply. This trend suggests that middle-range customer indices are most strongly associated with acquisition success, while extremely low or high indices may indicate prospects with lower likelihoods of acquisition. Employees: The PDP for employees shows fluctuations in the acquisition probability. A sharp rise is observed for firms with approximately 200–500 employees, peaking in this range. Beyond 500 employees, the probability declines, indicating that medium-sized firms are more likely to be acquired. This finding aligns with the understanding that mid-sized firms may present optimal opportunities for acquisition, balancing resource availability and potential for growth. Revenue: The plot for revenue illustrates a complex relationship with acquisition likelihood. The probability initially declines as revenue increases, reaching a low point, before rising and fluctuating for mid-range revenues. This indicates that acquisition likelihood is not directly proportional to revenue and may depend on other interactions, such as industry or firm size. Mid-range revenue firms appear to offer better acquisition opportunities. Key Takeaways: Non-linear Effects: The PDPs demonstrate non-linear relationships between the predictors and acquisition likelihood, highlighting the value of using the Random Forest model, which can capture such complexities. Customer Profiles: Mid-range customer indices, employee counts, and revenue levels are associated with higher acquisition probabilities, providing actionable insights for targeting acquisition efforts. Strategic Targeting: These findings suggest focusing on medium-sized firms and those with moderate revenue levels to optimize acquisition success. These results reinforce the Random Forest model’s ability to uncover nuanced relationships and guide effective customer acquisition strategies.

A graph of a customer

Description automatically generated



A graph showing the number of employees

Description automatically generated

A graph with lines and numbers

Description automatically generated

1. **Conclusions and Recommendations**

The Random Forest model consistently achieved the highest accuracy compared to other models, such as logistic regression and decision tree, according to the results from the QMD file. Its ability to handle complex, non-linear relationships among variables proved crucial for achieving superior predictive performance. The decision tree model, while interpretable and simple, underperformed due to its limited capacity to capture interactions and non-linear effects. Logistic regression, despite being a robust and interpretable model, also fell short in terms of accuracy and adaptability to non-linear data structures.

Variables like customer, employees, revenue, and acq\_exp\_sq were consistently identified as the most significant drivers of customer acquisition. The Random Forest model’s variable importance rankings highlighted these variables’ strong predictive power. Partial Dependence Plots (PDPs) revealed non-linear relationships for variables such as duration and profit, emphasizing the importance of using flexible models like Random Forest to uncover these insights.

The removal of variables that perfectly predicted acquisition ensured the integrity of the modeling process. However, the small number of remaining features limited the model’s complexity, particularly for logistic regression.

Our recommendations include first to adopt the Random Forest Model for Customer Acquisition. The Random Forest model should be used as the primary tool for predicting customer acquisition. Its accuracy, robustness, and ability to handle non-linear interactions make it the most effective choice for this task. This model is particularly suited for identifying high-value prospects, helping prioritize marketing resources efficiently.

Second, we should leverage Key Variables. Focus acquisition strategies on the most impactful variables, including customer, employees, revenue, and acq\_exp\_sq. These variables provide actionable insights for targeting potential customers and tailoring acquisition campaigns.

Third, we should refine Portfolio-Based Strategies: For high-value customers, optimize marketing efforts based on insights from the Random Forest model. For low-CLV customers or churn-prone groups, explore additional predictors or develop separate models tailored to these segments.

Fourth, we recommend to enhance Dataset Features: While the dataset was relatively clean, the inclusion of additional variables—such as customer engagement metrics or industry-specific features—could further improve the model’s predictive power. Incorporating external data sources may also enhance results.

Fifth, we recommend to future Modeling Efforts: Investigate advanced ensemble methods, such as Gradient Boosting or XGBoost, to compare performance with the Random Forest model. Conduct retention analysis to complement acquisition predictions, offering a more comprehensive view of customer dynamics. Periodic Evaluation and Model Updates: Continuously monitor model performance and retrain using updated datasets to account for shifts in customer behavior and market trends.

Overall, by utilizing the Random Forest model and focusing on high-impact variables, the organization can streamline its customer acquisition process, effectively allocate resources, and maximize profitability.

**Appendix:**

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph with a triangle

Description automatically generated**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A diagram of a company's company

Description automatically generated