Case 4 Customer Retention

Executive Summary

- A random forest model was identified as the optimal approach for predicting customer acquisition.
- For predicting the duration of customer retention, the random forest model achieved a highly accurate MAPE of 2.783%.
- Analysis of customer retention revealed that the most significant factors were crossbuy, followed by ret_exp and freq.

Problem

Our aim in this case study was to analyze the acquisitionRetention dataset to predict two key outcomes: will a customer be acquired and the expected duration of that the customer will be with us. By testing multiple model, we will be able to leverage the variables to make these predictions, evaluate variable importance, and optimize hyperparameters for better predictions of acquired customers. Additionally, we will compare the accuracy of the Random Forest model with a Decision Tree and a Logistic Regression model in predicting customer acquisition.

Lit. Review

The research we found evaluates ten machine-learning models, including Support Vector Machines (SVM) and Random Forest, to assess their effectiveness in predicting customer churn. The results show significant differences in performance among these models. Random Forest achieved the highest accuracy at approximately 96%, showing how good it is at handling complex datasets. SVM also performed well, with an accuracy rate of around 94%, making it a strong candidate for churn prediction. Simpler models like Logistic Regression showed lower accuracy, around 86%, suggesting limitations in capturing complex relationships within the data. These findings emphasize the importance of selecting appropriate models based on the complexity of the problem and the nature of the dataset.

Methods

Logistic

The logistic regression model was applied to predict customer acquisition, emphasizing interpretability and statistical rigor. To enhance the model's performance, we ensured the categorical data was balanced and applied backward selection to retain only the most significant

variables. The final model included acquisition expense, industry type, revenue, and the number of employees as predictors. Results revealed that industry type had the largest impact on acquisition likelihood, with customers in the B2B sector being 5.02 times more likely to be acquired than those in other industries. Revenue also played a significant role, with each additional million dollars in annual sales increasing the odds of acquisition by 7%, while larger employee counts also positively influenced acquisition likelihood.

The logistic regression model's accuracy reached 81.8%, just behind the tuned random forest model. Despite slightly lower accuracy, its advantage lies in the ease of interpretability, allowing for clear insights into the factors influencing acquisition. The logistic regression coefficients highlighted a "sweet spot" for acquisition expense, suggesting diminishing returns after a certain point, and provided actionable recommendations for targeting high-potential clients. These results underline the value of logistic regression as a complementary tool to more complex machine learning models, offering robust insights while remaining interpretable and practical for decision-making.

SVM

Our SVM model was trained using a radial basis function (RBF) kernel, with hyperparameters tuned over a range of values for gamma (0.01 to 0.1) and cost (0.1 to 1). The optimal parameters identified were gamma = 0.1 and cost = 0.6, which were used to train the final model. The model resulted in 123 support vectors, with a balanced representation from both classes.

Model performance was evaluated using a confusion matrix, which showed that the model correctly classified 76 observations (37 from class 0 and 39 from class 1) while misclassifying 22 instances. These results show that the model has a moderate ability to differentiate between the two classes. To try to improve the performance, we experimented with extending the hyperparameters and modifying gamma and cost values, but the changes did not result in significant performance improvements. We also tried changing the kernel, but we didn’t see any significant improvements to the model.

Random Forest

For acquisition, the tuned random forest model achieved an accuracy of 82.8%, outperforming logistic regression (81.8%) and decision trees (78.8%). Key predictors included industry type, revenue, number of employees, and acquisition expenses. Partial dependence plots revealed that customers in the B2B sector, with higher revenue and larger employee counts, were more likely to be acquired. A sweet spot in acquisition expenses ($400-$600) further maximized acquisition likelihood. The results underscore the importance of targeting B2B clients with optimized marketing strategies.

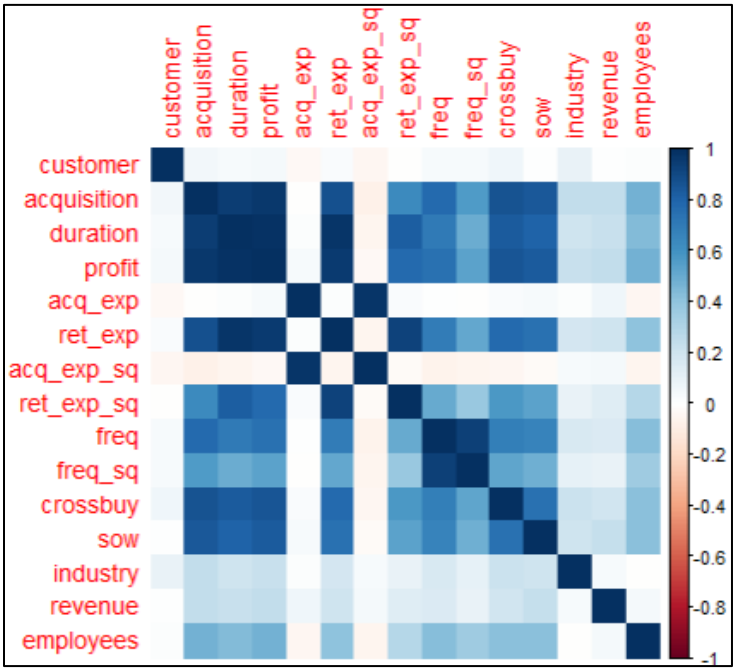
For duration prediction, the tuned random forest regression model showed strong alignment with actual values, with a significantly reduced mean squared error (MSE) after tuning. Retention expense emerged as the most influential predictor, positively correlating with longer customer durations, alongside variables like share of wallet and crossbuying behavior. For the logistic

model, we ensured the categorical data was balanced and applied backward selection to identify the most significant variables, enhancing the model's interpretability. Logistic regression results highlighted the importance of industry type, revenue, and employee count, providing clear actionable insights for targeting high-potential customers. By combining variable importance, interaction analysis, and partial dependence plots, the study highlights actionable metrics to enhance customer retention and acquisition while optimizing resource allocation.

Data

Our dataset included 15 variables and 500 observations, with no missing values or NA's. While we observed a number of zeroes, these were correlated with the acquisition column, which indicates whether a prospect was acquired. As part of the data cleaning process, we looked into the correlations (graph below) among these variables and removed those with the highest correlation to reduce multicollinearity. Also, we removed one of the customer ID fields. The variables removed from the dataset include customer, duration, profit, ret_exp, ret_exp_sq, freq, freq_sq, crossbuy, and sow. No additional data cleaning was necessary.

Correlation Plot



Results

Below is a comparison of the models evaluated, with the GLM model achieving the highest accuracy (0.7917), sensitivity (0.7917), and specificity (0.7917). However, the Random Forest model offers competitive performance with an accuracy of 0.7813, the highest sensitivity at 0.8000, and specificity of 0.7647. The higher sensitivity of Random Forest makes it particularly

effective at identifying true positives, which is critical for accurately predicting customer acquisitions. Additionally, Random Forest provides valuable insights through variable importance metrics, handles non-linear relationships effectively, and is robust to diverse datasets, making it a practical and interpretable choice for optimizing customer acquisition predictions.

Metric	Decision Tree	Random Forest	SVM	GLM
Accuracy	0.7188	0.7813	0.7760	0.7917
Sensitivity	0.7292	0.8000	0.7800	0.7917
Specificity	0.7083	0.7647	0.7710	0.7917

Retention

After predicting customer retention, we were able to build a highly accurate random forest model to predict the duration of customer retention, achieving a MAPE of 2.783%. This level of accuracy underscores the model's reliability. After building the model we dove into variable importance that revealed two distinct tiers: crossbuy emerged as the most influential variable, significantly outpacing the others, followed by ret_exp, ret_exp_sq, freq, and freq_sq, which also showed meaningful but comparatively lower impact.

	forest_D$importance
profit	25670.13591
acq_exp	725.52232
ret_exp	77222.00122
ret_exp_sq	67974.17551
freq	68541.14179
freq_sq	73248.47851
crossbuy	158233.01871
sow	27842.64825
industry	12.89682
revenue	540.56419
employees	1064.41807

Conclusion

The goal of this analysis was to predict customer retention and understand the factors influencing it, including the duration of retention. The GLM model performed best overall, achieving the highest accuracy (0.7917 each). However, the Random Forest model also performed well, with slightly lower accuracy (0.7813) but the highest sensitivity (0.8000), making it especially effective at identifying true positives in customer acquisitions. For predicting retention duration, the Random Forest model achieved a very accurate MAPE of 2.783%. Analyzing the variables showed that crossbuy was the most important factor, followed by ret_exp, ret_exp_sq, and freq with smaller but meaningful contributions. Alternative methods like Gradient Boosting (e.g., XGBoost) or Neural Networks could be explored in the future to potentially improve performance. Overall, the Random Forest model provides a strong balance of accuracy, interpretability, and insights, making it a valuable tool for predicting and optimizing customer retention.