Case 4 Customer Retention

Executive Summary

Our main problem is to use predictive modeling to predict which customers are likely to be acquired and how long they will stay. By solving this, we are trying to help businesses focus their resources on the right prospects and reduce marketing costs. This involves using the acquisitionRetention dataset to apply the models we have learned throughout this course, compare their performance, and identify the key factors for this case study.

ADD BULLETS ON KEY FINDINGS

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

add

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.

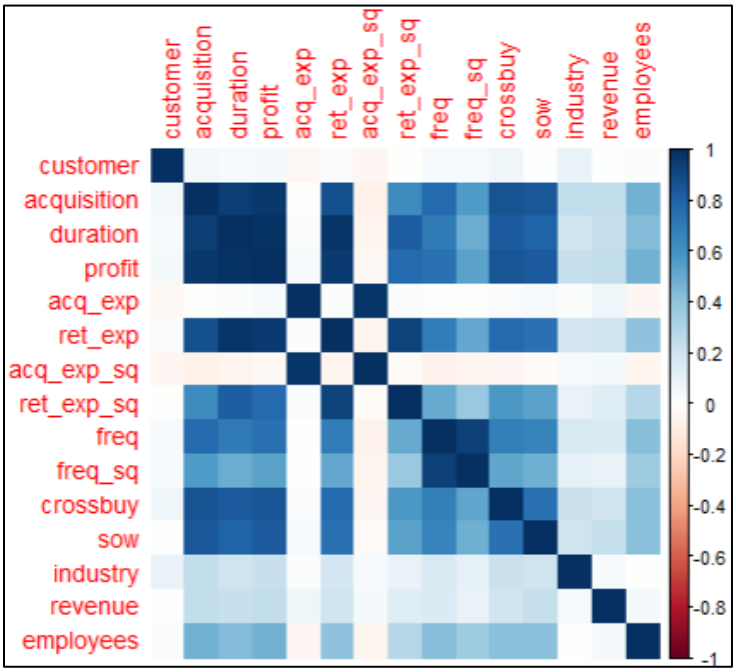
Random Forest

add

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 table comparing the accuracy, sensitivity, and specificity of the models evaluated. From the results, the Random Forest model demonstrates the highest performance across all metrics, with an accuracy of 0.7813, sensitivity of 0.8000, and specificity of 0.7647. The SVM model follows closely with an accuracy of 0.7760 and balanced sensitivity (0.7800) and specificity (0.7710). The Decision Tree model has the lowest performance, with an accuracy of 0.7188, sensitivity of 0.7292, and specificity of 0.7083. Overall, the Random Forest model is the most effective in predicting outcomes based on this dataset. UPDATE TO INCLUDE GLM

Conclusions

B

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