CASE STUDY #4

Customer Retention

DA 6813 DATA-ANALYTICS APPLICATION| 14-NOV-2021

KRISCHELLE JOYNER | KATHY KEEVAN | DAYANIRA MENDOZA | CHARLES REYES

Executive Summary

In the following case study, “Customer Retention”, we considered three different modeling methods to understand the factors that influence customer’s orders. The three different methods were linear regression, logistic model, and support vector machines.

After going through the results and comparing each model, we can see that using the linear regression model is not sufficient in this case study due to the non-linearity the models hold with the following data. This can be seen in the graph below. As for the following logistic models, we are able to predict the values of the categorical variables. Within the logistic modeling, we were able to find the S-curve by which we can classify the samples. For the SVM models, the results contained higher accuracy while containing all of the variables.

Background

The “SMCRM” package we will use in our analysis comes from the *SMCRM: Data Sets for Statistical Methods in Customer Relationship Management* book written by V. Kumar and J. Andrew Petersen. The book explores relationships between customer acquisition and customer retention to see if there are any correlations.

Managing customer retention and acquisition is essential for developing and maintaining customer relationships. The first step to cure customer retention and acquisition is to predict which customers have a high probability of ending their relationship with the firm and the probability of acquiring a new customer. The second step is to target the predicted at-risk current customers or new customers with high likelihood of joining using incentives such as pricing offers or communications such as emails. Models that accurately predict customer retention and acquisition are pivotal in targeting the right customers, thereby decreasing the cost of the marketing campaign and using scarce firm resources more efficiently.

The Problem

A balanced approach to customer acquisition and retention will help a firm achieve greater profitability by balancing the two processes. Firms that are able to understand the drivers of customer acquisition, customer retention, and customer profitability can then run simulation exercises with these results to increase profitability.

The analysis will primarily focus on:

Which variable is the most significant indicator of customer acquisition?

Of the 15 variables in our dataset, we want to find the best indicator of customer retention to understand customer behavior and how to change a firm's behavior to match those needs.

What is the expected lifetime of customers?

Having acquired a customer, we might be interested in knowing how likely it is that they will stick with us. In other words, how long do we expect to retain them?

Review of Related Literature

The **acquisitionRetention** dataset is part of the SMCRM library. SMCRM refers to the book *SMCRM: Data Sets for Statistical Methods in Customer Relationship Management* (Kumar and Petersen, 2012). The **acquisitionRetention** data is covered in Chapter 5. The definitions of the data fields can be found in the Appendix.

In Chapter 5 of the SMCRM book, the authors use the SMCRM data to (1) predict whether a customer will be acquired, and (2) if the customer was acquired, predict how long they would remain a customer (duration). They proposed the following modeling framework:

Acquisition model

$$z_i^* = \alpha' v_i + \mu_i \quad \text{(acquisition equation)}$$
$$z_i = \begin{cases} 1 & \text{if } z_i^* > 0 \\ 0 & \text{if } z_i^* \leq 0 \end{cases}$$

Where:

- z^*_i is a latent variable indicating customer *i*'s utility to engage in a relationship with a firm
- α is a vector of parameters
- v_i is a vector of covariates affecting the acquisition of customer *i*
- μ_i is an error term
- z_i is an indicator variable showing whether customer *i* is acquired ($z_i = 1$) or not ($z_i = 0$)

Kumar and Petersen chose what they felt would be the most likely predictors of acquisition and used a probit regression. All predictors were significant. They used both *Acq_Exp* and *Acq_Exp_SQ*. Below, Acq_Exp has a positive effect on acquisition, while the square of *Acq_Exp* has a negative effect on acquisition. This demonstrates a diminishing "bang for the buck": each additional dollar spent has less of an effect on whether a customer is acquired. B2B businesses (*Industry* = 1) are more likely to be acquired, as well as companies with higher revenue and more employees.

Variable	Estimate	Standard error	*p*-value
Intercept	−8.255	0.811	< 0.0001
Acq_Exp	0.017	0.002	< 0.0001
Acq_Exp_SQ	−0.000 02	0.000 002	< 0.0001
Industry	1.217	0.168	< 0.0001
Revenue	0.043	0.008	< 0.0001
Employees	0.004	0.0004	< 0.0001

The probit model output was used to calculate the model's predictive power. The predicted probability of repurchasing was calculated, and customers were split into two groups for predicted to adopt (probability >= 0.5) or not predicted to adopt.

		Actual acquisition		
		0	1	*Total*
Predicted acquisition	0	*105*	*38*	***143***
	1	*57*	*300*	***357***
	Total	***162***	***338***	***500***

The accuracy is 81% (405 / 500). The authors make the distinction of how accurately the model predicts who will adopt (88.7%) and who will not adopt (64.8%). The prediction of non-adopters is not as good as the prediction of adopters.

Duration model

$$y_{Di}\begin{cases} = \beta_D' x_{Di} + \varepsilon_{Di} & \text{if } z_i = 1 \\ = 0 & \text{otherwise} \end{cases} \quad \text{(duration equation)}$$

Where:

y_{Di} is the duration of customer *i*'s relationship with the firm
β_D is a vector of parameters
x_{Di} a vector of covariates affecting the duration of customer *i*'s relationship with the firm
ε_{Di} is an error term
z_i is an indicator variable showing whether customer *i* is acquired (z_i = 1) or not (z_i = 0)

Here the authors look at how long a customer is likely to stick around, given that adoption occurred (*Acquisition* == 1). More precisely:

$$E(Duration) = P(Acquisition = 1) * E(Duration | Acquisition = 1)$$

The authors introduce an additional variable from the acquisition model, the inverse Mills ratio, which is represented by lambda (λ). Lambda represents correlation in the error structure across the equations for acquisition and duration. Again, the authors choose variables they believe will be predictive.

Variable	Estimate	Standard error	*p*-value
Intercept	91.008	9.756	< 0.0001
Ret_Exp	2.528	0.029	< 0.0001
Ret_Exp_SQ	−0.001	0.000 02	< 0.0001
Freq	7.072	0.806	< 0.0001
Freq_SQ	−0.842	0.040	< 0.0001
Crossbuy	3.196	0.479	< 0.0001
SOW	0.353	0.045	< 0.0001
Lambda(λ)	29.520	2.557	< 0.0001

All predictors are significant. Again we see an inverse relationship of the squared expense variables to retention, indicating decreasing utility with each dollar spent.

Model predictive power was measured using mean absolute deviation (MAD):

$$MAD = Mean\{Absolute\ Value[E(Duration) - Duration]\}$$

The MAD for all customers = 144.02, which means the predictions of *Duration* deviate from the actual *Duration* by ~144 days on average. When we compare this to the mean *Duration* of 484.09 days, this is a significant improvement in predicting how long a company will remain a customer. /

(Kumar and Petersen, 2012)

Methodology and Assumptions

To find the best model, we will be training our models using 3 different techniques.

1. Random Forest
2. Decision Tree
3. Logistic Regression

Random Forest

The random forest model is a culmination of a large number of decision trees created using bootstrapping random sampling techniques to predict upon a response variable. The trees built in the random forest model are uncorrelated and use only a subset of predictors to choose from at each node and tree. They also work very well with data that has outliers, missing values, or is unbalanced. Using two random forest models: one with hyperparameters and one without, we will see which model generates the best results.

Decision Tree

Decision trees typically tend to overfit, but they are easy to understand. Also, they are pretty good at prediction and run efficiently. The two types of decision trees are classification and regression; because the predicted values are not continuous, we will run classification trees.

Logistic Regression

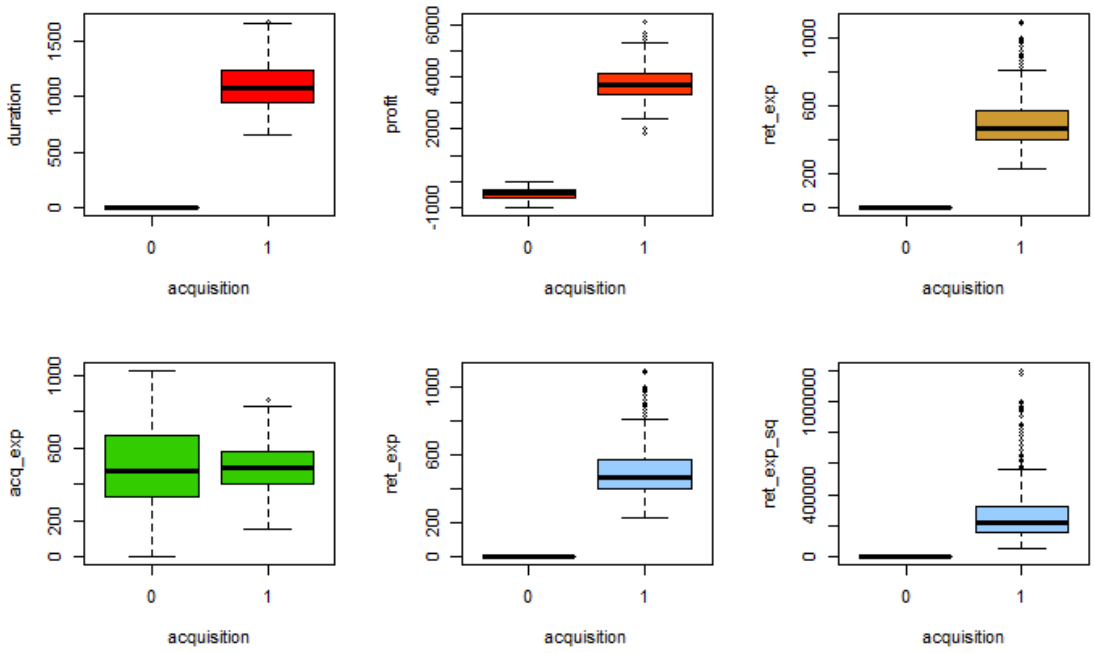
Logistic regression might be a good model for our data given it is easy to implement, interpret, and efficient to train. It makes no assumptions about distributions of classes in feature space. We will run a simple logistic model to see which model will yield better results.

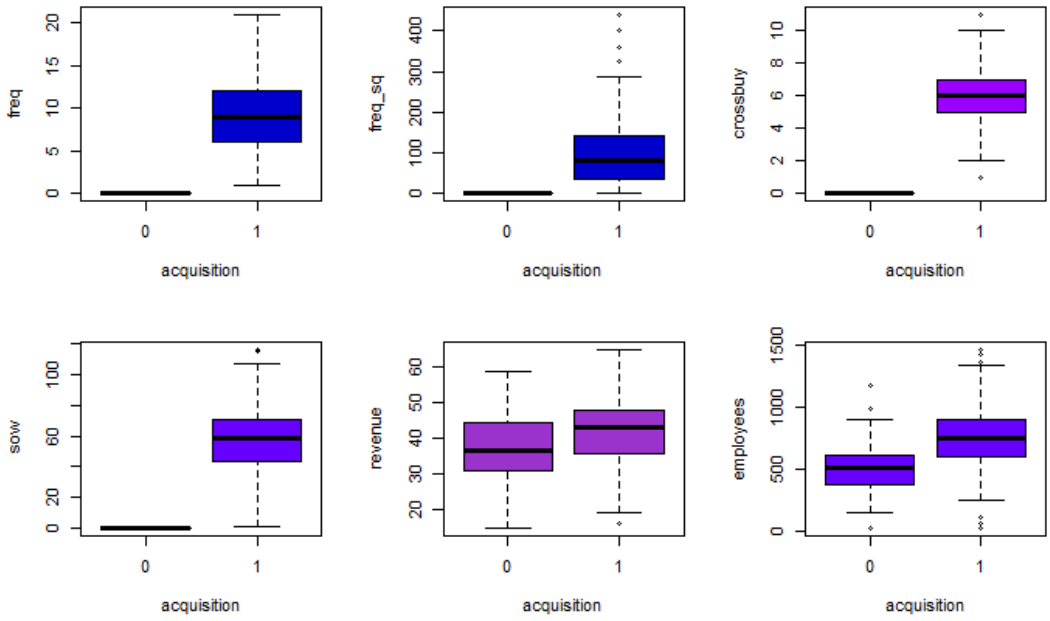
Data

The data set used for our modeling and predictions was sourced from the built-in “acquisitionRetention” data set in the SMCRM library in R. The data set contains 500 observations and 15 variables (see appendix for variable descriptions). The data set was clean and did not contain any missing values. All variables by default were imported as numeric data types. However, the acquisition and industry variables had binary values so they were converted to factor data types for our analysis.

For the first portion of our analysis, we used *acquisition* as our dependent variable. Knowing the *customer* variable is only an ID for each observation, we decided to exclude it from our model. For the remaining variables, we created boxplots to show the distribution of key variables. Many of these

variables only pertain to current customers (see the data elements in the Appendix), and for prospects the value is zero. For those variables that pertain to both current and potential customers, we can see that some variables, like *acq_exp* and *acq_exp_sq*, show a weak relationship between customers and non-customers. For the two acquisition expense variables, the median is almost equal and the interquartile range is tight. This shows that the company has spent about the same amount per customer on acquisition. On the other hand, *profit* shows a clear relationship between current and prospective customers. This makes perfect sense since current customers contribute revenue while prospective customers just contribute to expenses (i.e., negative revenue). We also see a relationship between customers and prospects in the *employees* boxplot; companies with more employees are more likely to be acquired as customers.





We included in the model only those variables that apply to both current and prospective customers. We excluded *acq_exp_sq* as it is merely a square of *acq_exp* and we do not believe it to be particularly useful as a predictor for client retention. This leaves us with model variables *acq_exp, industry, revenue,* and *employees*."

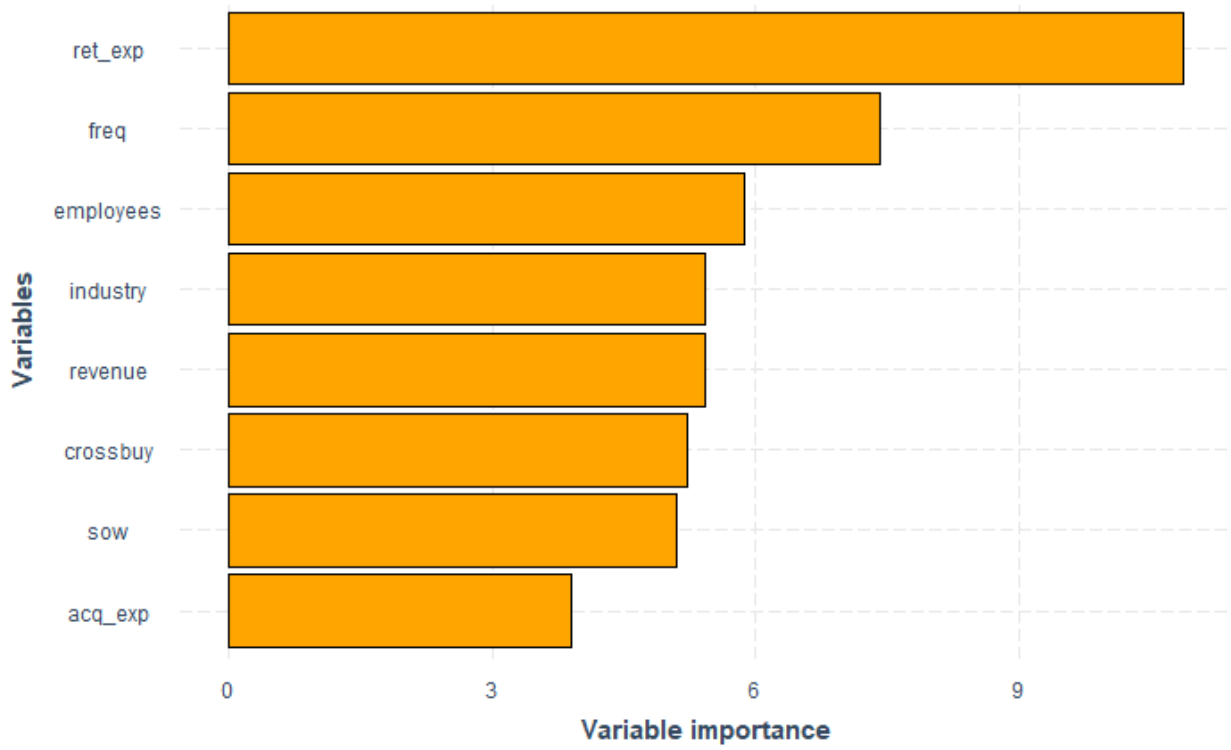
For the second part of our analysis, we used *duration* as our dependent variable. Similar to what we did for the first portion of our analysis, we excluded the *customer* variable. Leveraging the results from the first portion of our analysis as well as the original data set, we excluded all observations where acquisition was equal to 0. We also checked for any multicollinearity issues and noticed some potential concerns with the *profit, acq_exp*, and *ret_exp* variables. Once we removed the highest variance inflation factor, it relieved the concerns for the remaining variables. The following variables were used in our final model: *acq_exp, ret_exp, freq, crossbuy, sow, industry, revenue*, and *employees*.

For both models, the data set was split into train and test data sets using an 80% and 20% split, respectively.

Findings

Our analysis showed that the tuned random forest model had the highest accuracy rate of all the models tested. It produced an accuracy rate of 82.8%. The hyper parameters for this model were 1000 trees and an mtry value 2. The logistic model was the next model highest at 81.8%, with the decision tree following at 78.7%.

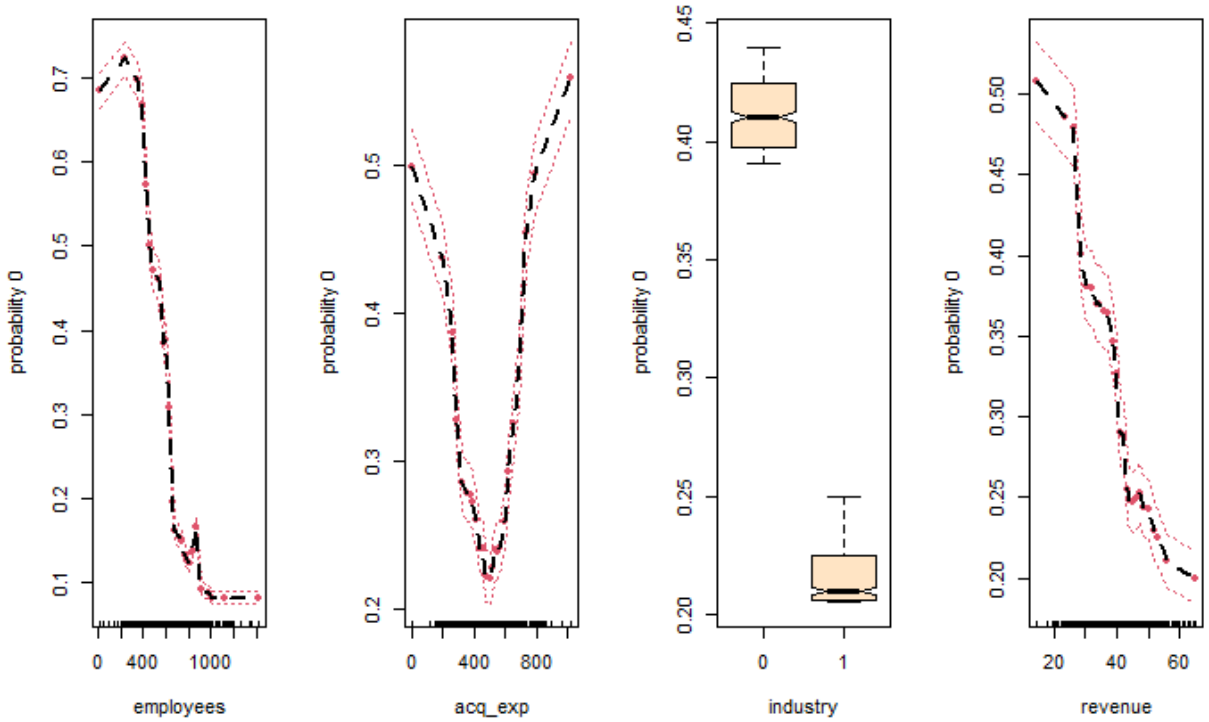
In the analysis we computed the variable importance to detect interactions and optimize hyperparameters for acquired customers. By doing so, we notice that the variables were negative. We then added a large constant to them and then took the log and plotted the results.



The decision tree is a non-parametric test that has no assumptions. It simply splits the data into nodes that can be used to classify and predict on the data. The nodes are determined by which split in the data or variable would reduce the residual sum of squares the most and thus the split is made at this point.

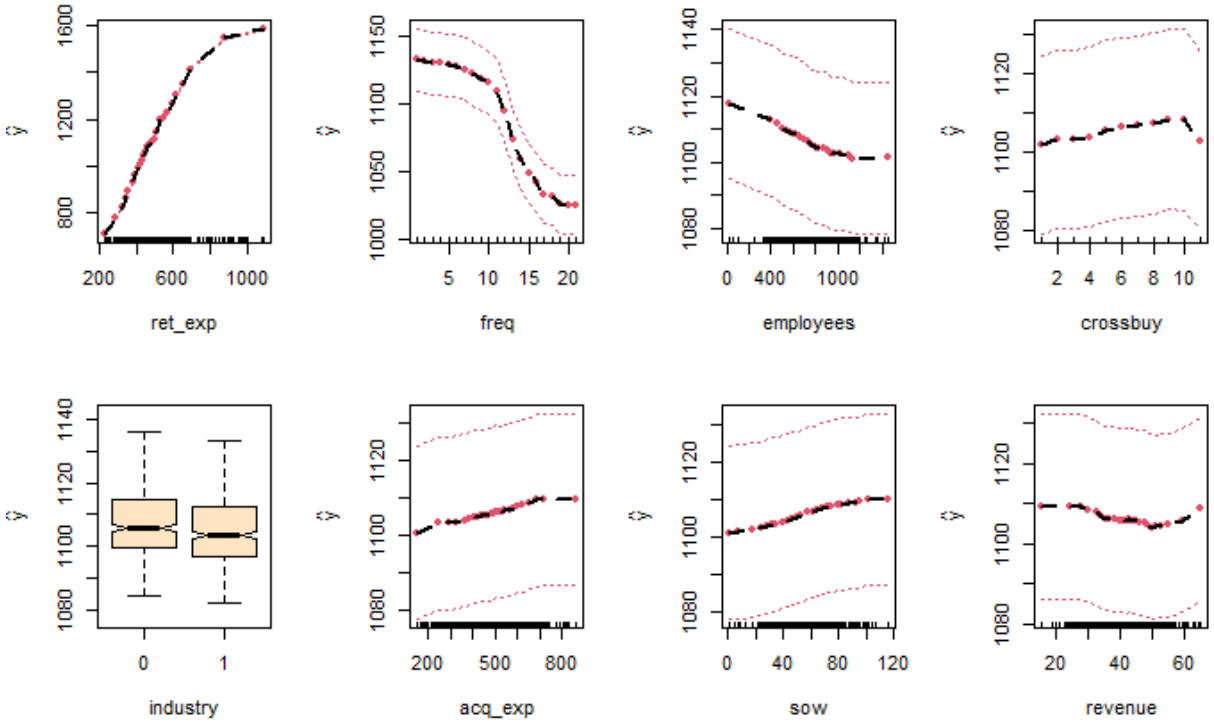
In the logistic model, under the assumption that that contributes to the accuracy of the results of the model. If we summarize the results of the logistic model, the p-value of acq_exp is above the selection threshold, so we fail to reject the null hypothesis. There is no significant statistical evidence that changes in acq_exp increase the odds of gaining an acquisition, holding all other variables constant. The p-value for industry is below the threshold, so we reject the null hypothesis that the coefficient is equal to zero. The odds of an acquisition are e^1.614 (5.02) times higher for a customer who is in the B2B industry compared to one who is not, when all other variables are held constant. The p-value for revenue is below the threshold, so we reject the null hypothesis that its coefficient is equal to zero. The odds of an acquisition increase by a factor of e^0.07 (1.07) for each million in annual sales revenue, holding all other variables constant. The p-value for employees is below the threshold, so we reject the null hypothesis that it's each additional employee in the prospect's firm, holding all other variables constant.

Partial Dependency Plots



Partial dependence plots illustrate how each variable affects the model's prediction. The idea behind how they are interpreted is similar to the interpretation of coefficients in a linear regression model.

For predicted *acquisition*, we see that there is an inverse relationship between the probability of no acquisition and number of employees and annual sales revenue of the prospect's firm in millions of dollars. Another way to look at it would be: holding all other variables constant, the more employees in a prospect's firm, the higher the likelihood of acquisition. The same holds true for annual sales revenue of the prospect's firm. We can also see that there is a "sweet spot" for *acq_exp* that maximizes the probability of *acquisition*: somewhere between 400 and 600 dollars. We also see that the probability of *acquisition* is highest when the customer is in the B2B industry.



Here we have the partial dependence plots for predicting *duration*. There is a direct relationship between predicted *duration* and *ret_exp, acq_exp* and *sow*. There is a generally inverse relation between *duration* and *freq* and *employees*. Predicted *duration* generally decreases as *revenue* increases, until about the 50-million dollar mark, at which point it begins to increase again. We also see an interesting dip between predicted *duration* and *crossbuy*, where duration generally increases until a *crossbuy* of about 10 product categories, at which point the predicted *duration* begins to drop off. It is difficult to determine whether there is a significant difference between *industry* and predicted *duration*, as there is a lot of overlap in predicted *duration* when the customer is in the B2B industry, or when the customer is not in the industry.

Conclusion and Recommendations

In conclusion, we recommend using the tuned random forest model for prediction of customer acquisition. It is the most accurate model, and it accounts for interaction and non-linearity of the variables in its methodology, versus the logistic model in which we need to account for them manually which is not always easy to do. Acq_expense and employees were the most important and significant predictors of the response variable, acquisition.

We believe it would also benefit our study if we were to analyze what variables and models have the greatest prediction power of how long a customer stays with the company. The ability to identify and accurately predict the acquisition of a customer as well as the lifetime of that customer would be of great value to a company. In doing this it would allow the use of many of the variables that we were

required to cut from the model of acquisition. The tune random forest model is a relatively new methodology and will surely see many improvements and new potential uses in the future.

Bibliography

Kumar, V., Petersen, J.A., 2012. Balancing Acquisition and Retention, in: Statistical Methods in Customer Relationship Management. John Wiley & Sons, Ltd, pp. 121–148. https://doi.org/10.1002/9781118349212.ch5

Appendix

Format

Data frame with the following 15 variables:

customer	customer number (from 1 to 500)
acquisition	1 if the prospect was acquired, 0 otherwise
duration	number of days the customer was a customer of the firm, 0 if acquisition == 0
profit	customer lifetime value (CLV) of a given customer, -(Acq_Exp) if the customer is not acquired
acq_exp	total dollars spent on trying to acquire this prospect
ret_exp	total dollars spent on trying to retain this customer
acq_exp_sq	square of the total dollars spent on trying to acquire this prospect
ret_ext_sq	square of the total dollars spent on trying to retain this customer
freq	number of purchases the customer made during that customer's lifetime with the firm, 0 if acquisition == 0
freq_sq	square of the number of purchases the customer made during that customer's lifetime with the firm
crossbuy	number of product categories the customer purchased from during that customer's lifetime with the firm, 0 if acquisition = 0
sow	Share-of-Wallet; percentage of purchases the customer makes from the given firm given the total amount of purchases across all firms in that category
industry	1 if the customer is in the B2B industry, 0 otherwise
revenue	annual sales revenue of the prospect's firm (in millions of dollar
employees	number of employees in the prospect's firm

Source: [acquisitionRetention: Acquisition-Retention Data from Chapter 5 in SMCRM: Data Sets for Statistical Methods in Customer Relationship Management by Kumar and Petersen (2012). (rdrr.io)]