Customer Retention Case Study

Justin Howard, Lily He, Sadiyah Hotakey, Michael Adebisi, Olamiposi Sunmola

Executive Summary

This case study aims to predict which customers will be acquired and for how long (duration) based on a feature set using a random forest, compute variable importance to detect interactions and optimize hyperparameters for acquired customers, as well as compare the accuracy of model with decision trees and logistic regression model for acquiring customers. <Insert result interpretations>.

Problem

Managing customer retention and acquisition is essential for developing and maintaining customer relationships. The first step to curing 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 a 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.

Review of Related Literature

When dealing with classification, some popular methods/models are Logistic Regression (LogitR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest (RF) and Support Vector machine (SVM). All of these methods can handle the special case of a binary variable. While some of them provide higher simplicity, interpretability and robustness (LogitR, LDA, QDA), the others tend to perform better in the most complex scenarios with difficult relationships between inputs and outputs (RF, SVM). Nonetheless, all of them are considered valid choices for almost any binary-output setup, and are often compared before picking the final model. For the sake of exposition, in this case study we use the LogitR model, which in addition to the advantages mentioned above, presents very few impositions in terms of assumptions to verify. To do so, in R we rely on the glm() function of the stats package. We use the sensitivity, specificity and accuracy, which provide an indication of the ability of the model to accurately predict whether the customer will purchase the book or not. For this purpose, one of the functions we relied on was the confusionMatrix() function of the caret package.

Methodology

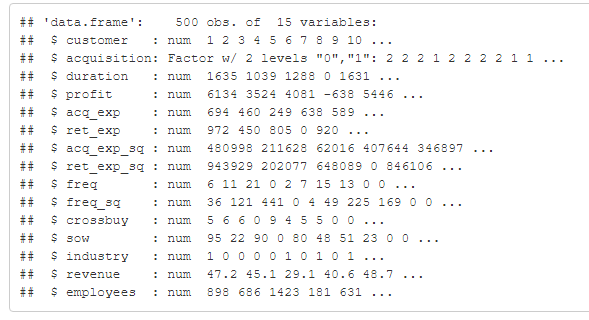
There are three modeling techniques that will be used in our analysis. Random Forest, Decision Tree, and Logistic Regression will be the techniques used.

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the stock prices and evaluates risks of stocks from the data features. A tree can be seen as a piecewise constant approximation.

Logistic regression is a modeling technique that assumes a linear relationship between the independent variables and the response variable’s log odds. The response variable is to be categorical without multicollinearity among predictors. A third assumption is that the observations are independent and distributed identically. Limitations include that it constructs linear boundaries; assumption of linearity between the dependent variable and the independent variables; can only be used to predict discrete functions; cannot solve non-linear problems due to its linear decision surface.

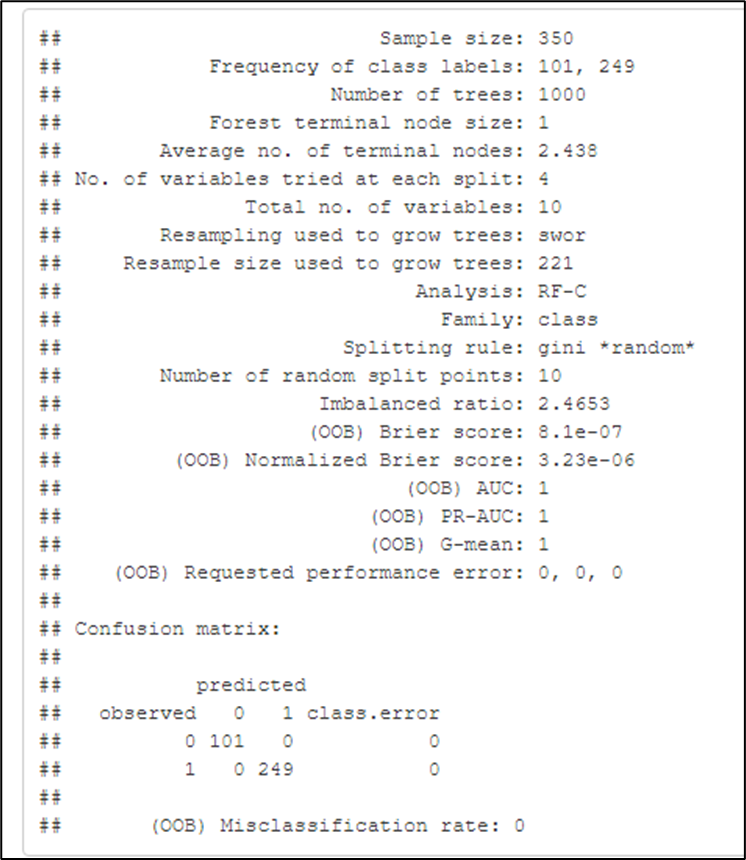
The Data

For this case analysis, we used the acquisitionRetention dataset in the SMCRM package. This data set contains 500 observations and 15 variables. There are no missing values in this dataset. The variables were then converted to factor variables. After which, we split the data into test and training datasets.

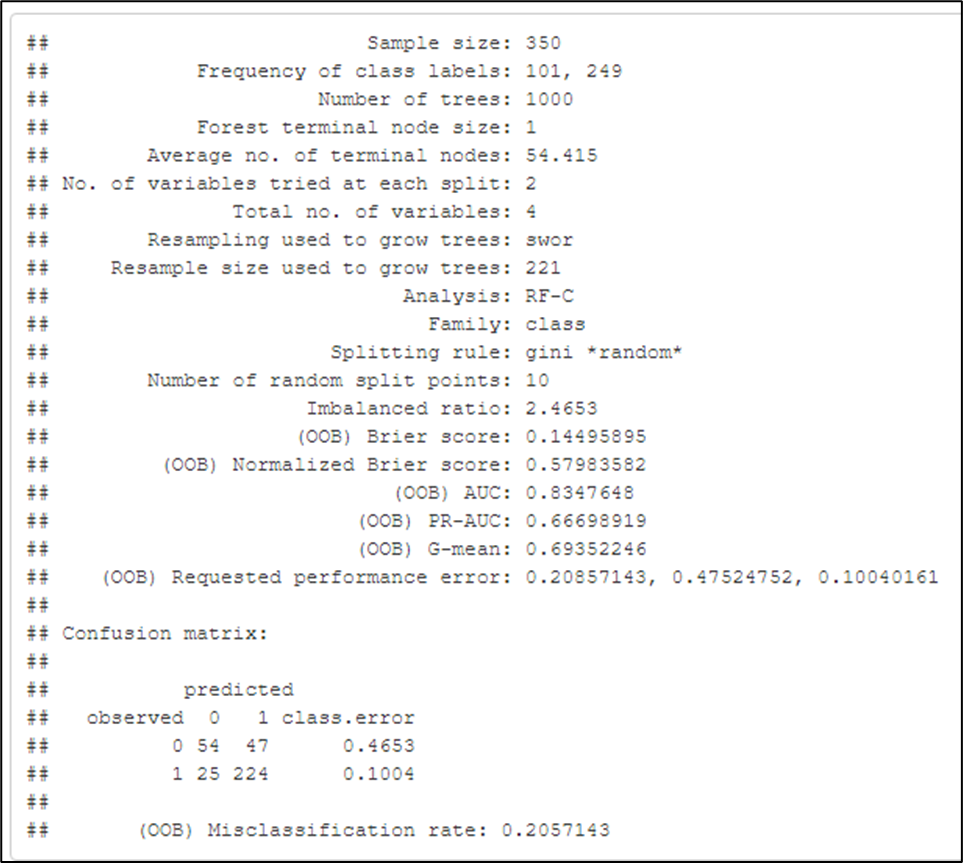


Findings

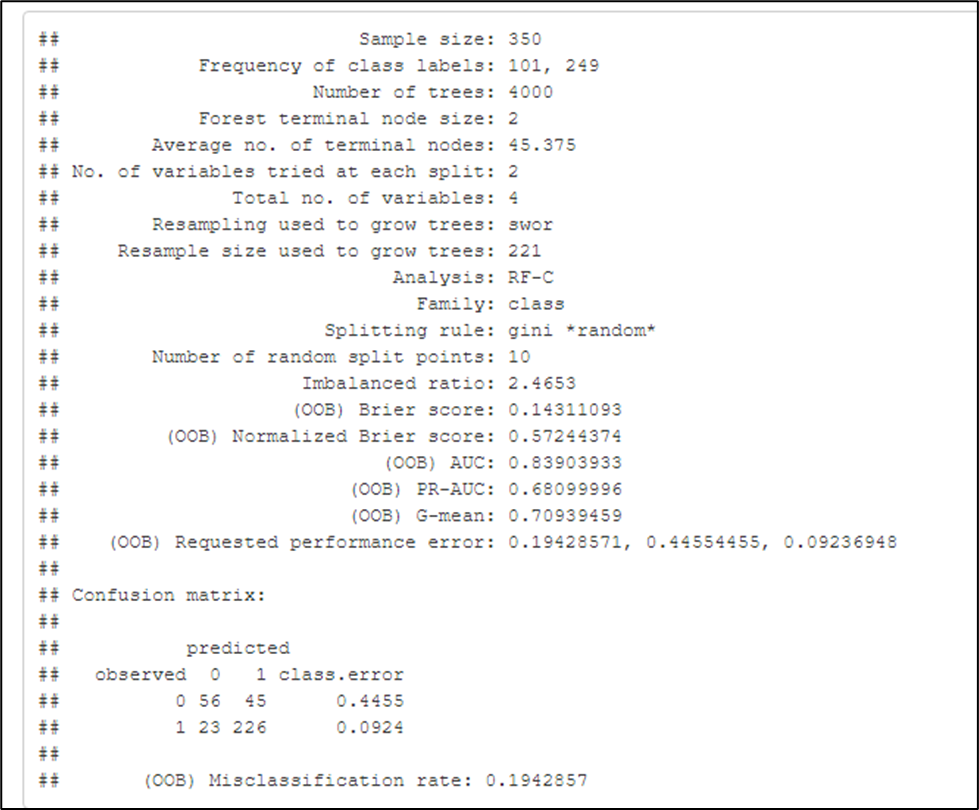
Initially the SMCRM and random forest SRC packages were installed to initiate the dataset. The dataset was reviewed for missing values, and it was found that there are no missing values in the dataset. The train test split is run to estimate the performance of Machine learning Algorithms and it was found that ML algorithms are run perfectly i.e., no error found. Then the random forest algorithm was run using acquisition as Dependent variable and rest of the variables as the independent variables. Accuracy was found to be 100% with 0 misclassification rate.



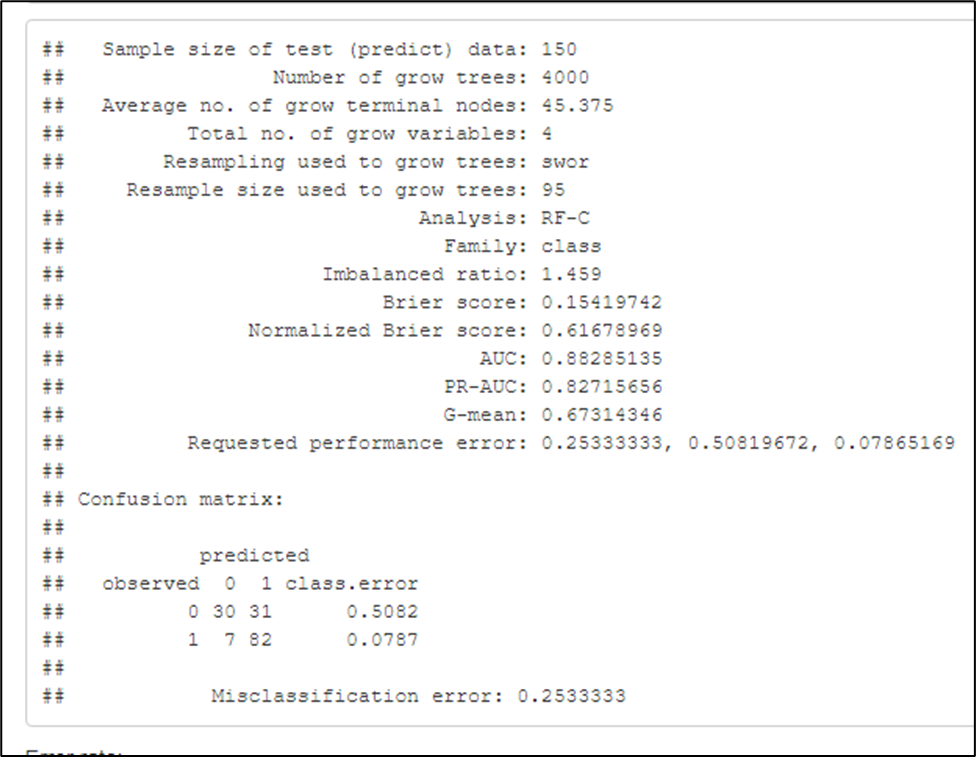
Then the linear and non-linear parametric models were used based on binary regression models with different number of independent variables. This time accuracy of the model reduced to 79.43% with a misclassification rate of 20.57%.



Then the hyper-parameters were used to train the random forest model so that prediction can be made, and 4,000 trees were considered. With hyper-parameters misclassification rate was found to be 19.43% and accuracy was 80.57%.

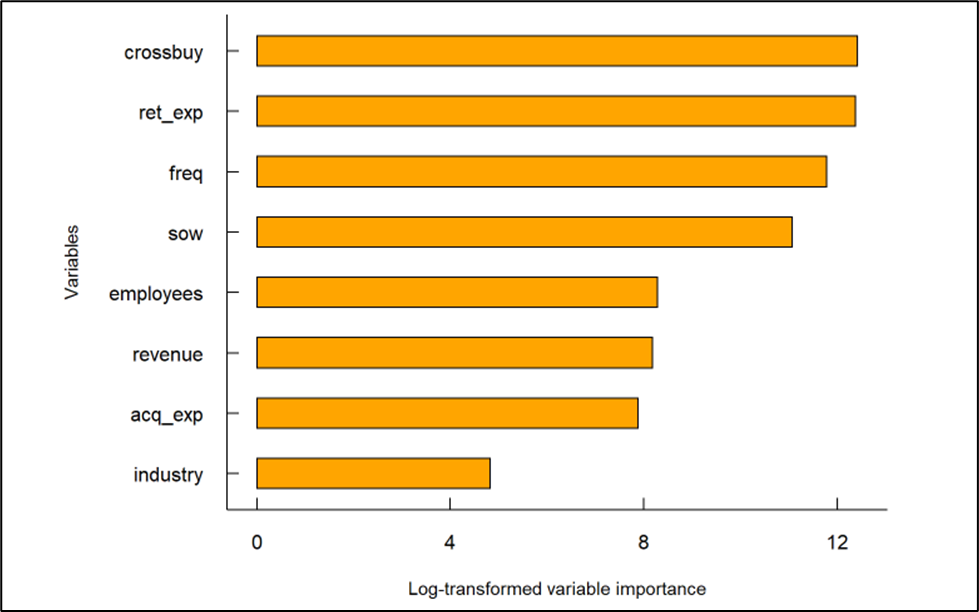


With no optimal hyper parameters, the misclassification rate was 25.33% and accuracy was 74.67%.

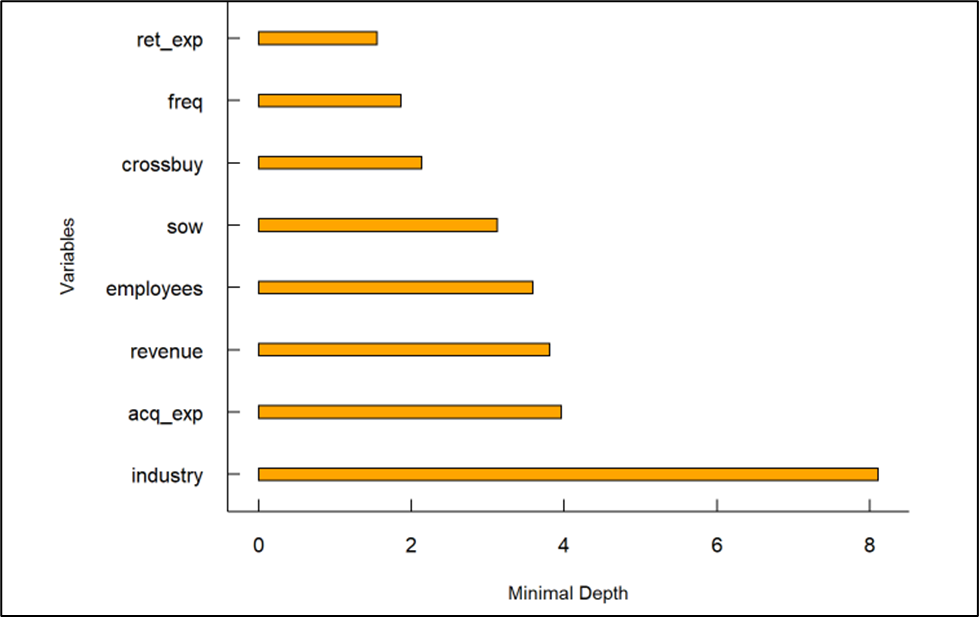


Forest\_acqhyper is slightly better than forest\_acq3 Forest\_acq3 error (on training set): 79.41 Forest\_acqhyper error (on training set): 80.57 Forest\_acqhyper error (on test set): 74.6667 dt\_fit (decision tree not pruned on test set): 75.33 dt\_pfit (decision tree pruned on test set): 80.67. Now a new variable NEW\_df was added to the dataset. The problem was to predict classification (acquisition) and then use accurate predictions to train a regression model for duration. Random forest regression was run and for all the variables being added it gave a R squared value of 0.9925 so it means 99.25% variation in dependent variable is being explained by all of the independent variables. Duration Forest inference Variable importance (five different tools on relationship between DV and IVs) was found. We got the values for their part involved in the action of the dependent variable.

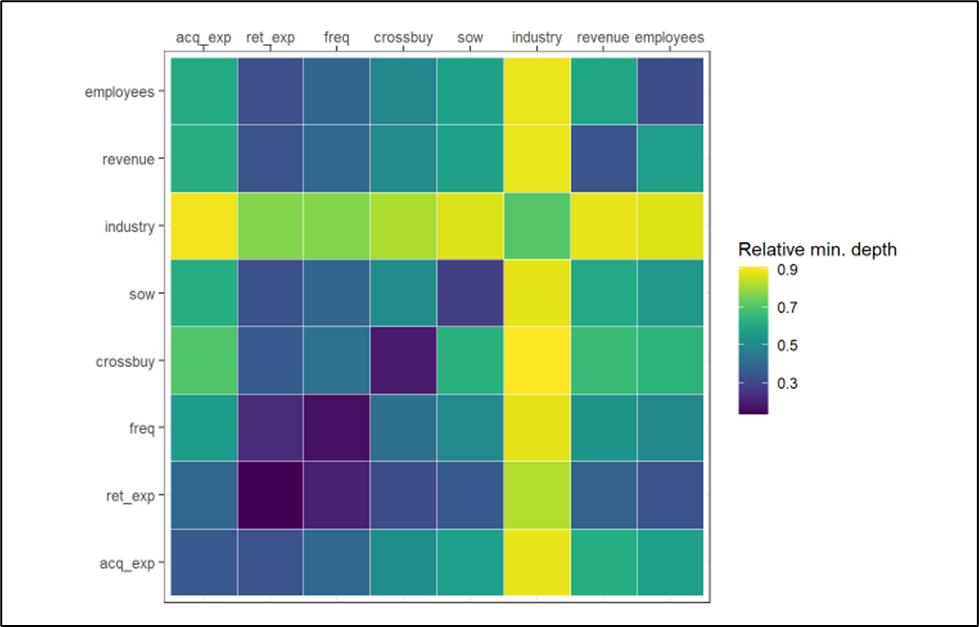
A graph was drawn to visualize the variable importance for all the variables. It was found that crossbuy, ret\_exp, freq and sow are significantly important. The rest of the variables can be ignored for better accuracy of the model. But it was getting difficult to assess the variable importance for the remaining important variables i.e., employees, revenue, acq-exp and industry. For this log transformation method was used and it was found that crossbuy had the highest value of 12 while the least was with the industry i.e., 4.



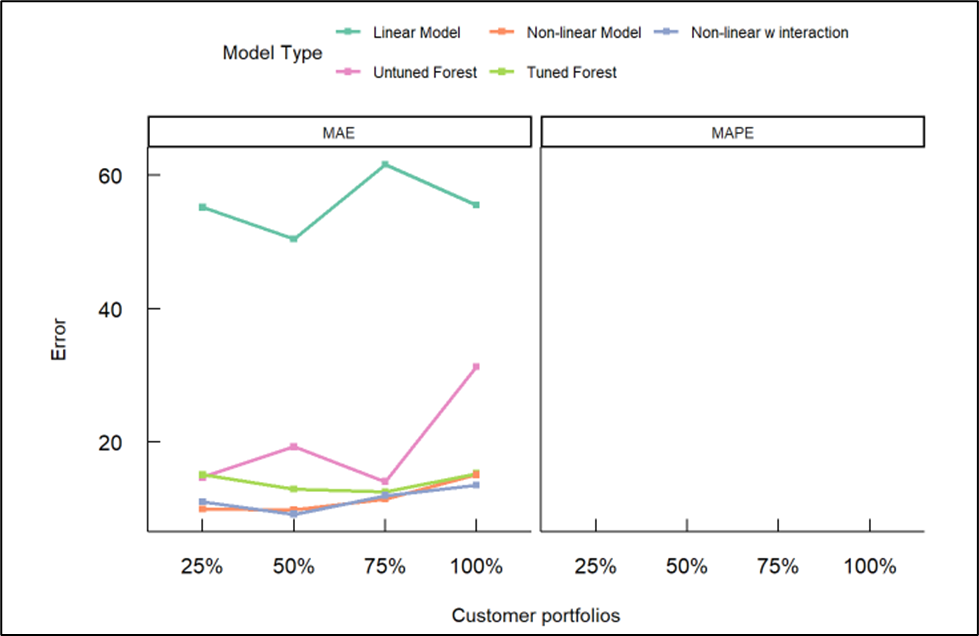
Now minimal depth graph was used to visualize the variable’s depth. It was found that industry has the more minimal depth value of 8 and crossbuy had the least value i.e., 2. The smaller the value of the minimal depth the more the variable will be.



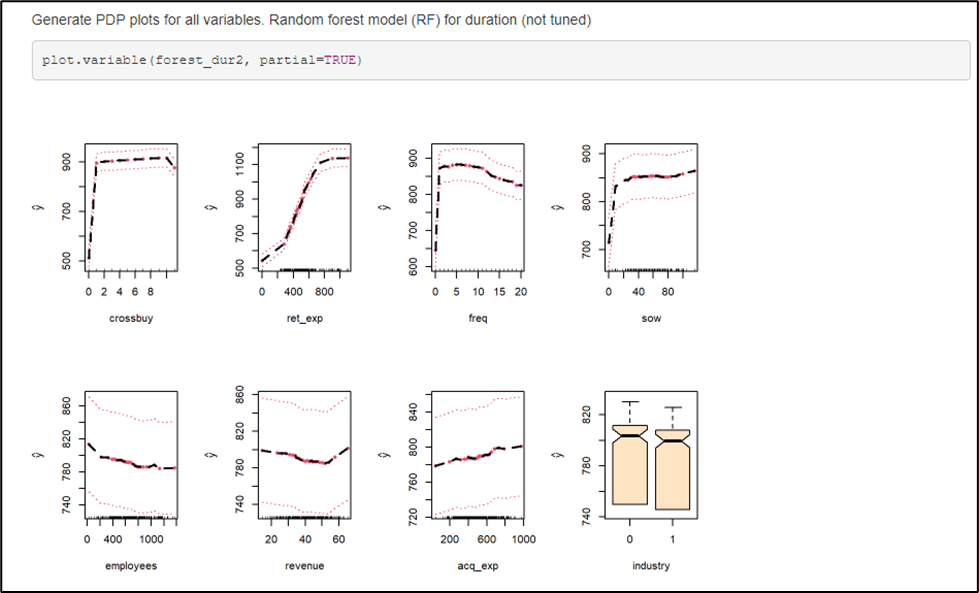
Then using the relative minimal depth value, a colorful matrix was drawn to show the variable interaction. Employees and industry are interacting in the yellow shade which means they are strongly related to each other, or we can say dependent on each other likewise revenue and industry, acq\_exp and industry.

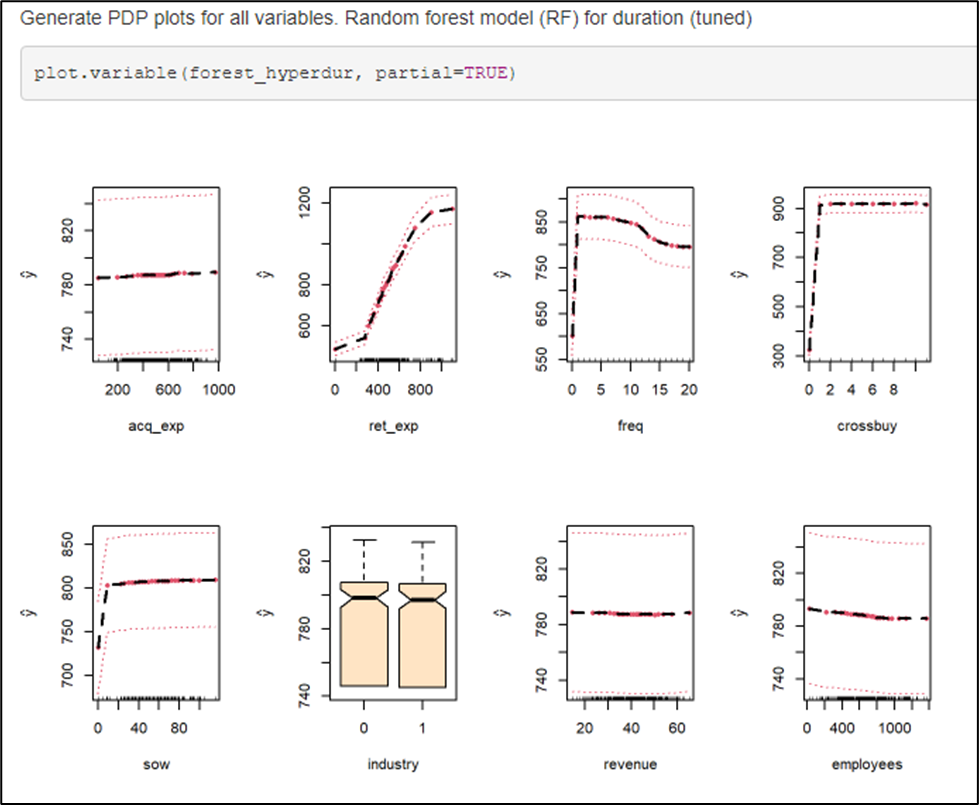


A linear regression model was run using cus\_ret as the dependent variable with the rest of the as independent variables. For this model, we got the R squared value of 0.9784 which means 97.84% of the variation is being explained by the independent variables. Then the nonlinear variables were added to the model and regression was run again. Now the R squared value increased to 0.9999 which means we got a 99.99% best fit model for making predictions. To compare the models a graph was drawn to see which model was the best with lowest error rate. The nonlinear model and nonlinear vs interaction models provided the least error rate with a 100% customer portfolio.



PDP plots were run to estimate the performance of each variable in the model. Pruned decision tree was run to predict customer acquisition and the accuracy of the decision was found to be 80.67%.





Conclusion

We have found the non-linear regression model with 99.99% accuracy, the random forest test gave us 0 misclassification showing 100% accuracy/performance, with the relative minimal depth matrix we found that industry has strong relations/interaction with employees, revenue and crossbuy. In the end, we got the pruned decision tree with 80.67% accuracy.

References

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), 352-359.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. In *Robust data mining* (pp. 27-33). Springer, New York, NY.

Bose, S., Pal, A., SahaRay, R., & Nayak, J. (2015). Generalized quadratic discriminant analysis. *Pattern Recognition*, *48*(8), 2676-2684.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157-175). Springer, Boston, MA.

Team, R. C., Team, M. R. C., Suggests, M. A. S. S., & Matrix, S. (2018). Package “Stats.”. *The R Stats Package*.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Team, R. C. (2020). Package ‘caret’. *The R Journal*, *223*, 7.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

Vapnik V., ”Statistical Learning Theory”, Wiley, New York, 1998.

Loh, Wei-Yin. (2011). Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 1. 14 - 23. 10.1002/widm.8.

Appendix

R code attached separately.