BookBinders Case Study

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

Executive Summary

This case study aims to address which customers Bookbinders should target and how much more profit the company would generate if data analytic models are used as opposed to sending the mail offer to an entire list of customers. Data analytic models being considered are linear regression, a logit model, and support vector machines to develop a highly accurate model. The output we hope to have would be categorical. Looking at the linear regression, we concluded that it is not appropriate because the diagnostic plots are not easily interpreted since this is a classification problem. In the logit model, we wanted to evaluate the predictive performance on optimized and un-optimized models on our training and test datasets. For the unoptimized model confusion matrix displays accuracy of 89.74% while the optimized model has a 70% accuracy. In addition, three different support vector machines were run on this data, in which we found the polynomial support vector machine has a high accuracy of 95.3%, highest of all three support vector machine models but indicates less sales of the book. While the linear and radial SVM would save a good amount in mailing costs.

Problem

This case study considers a book club company, Bookbinders Book Club (BBBC) that sells specialty books via direct marketing. BBBC has a large database and sends out mailings monthly. Performing a naïve campaign in which there is not a targeted audience will prove costly and inefficient. The company decides to utilize statistical techniques to better target their audience. They wish to explore three predictive modeling techniques and determine which model is the most accurate so that they can identify which customers to target in their upcoming mail campaign in the Midwest. In this study, we will utilize different methods to identify customers most likely to purchase a book as well as identify which attributes are the most influential when it comes to predicting customer purchases. We predict the SVM, or the logit model would be the best model to fit the data. We will then do a cost and market analysis to determine which model will be the most profitable.

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. Multiple Linear Regression (MLR), logistic regression, and support vector machines (SVM).

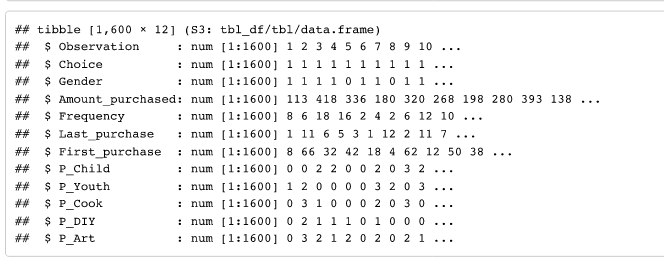
Linear regression attempts to model the relationship between an explanatory variable and a dependent variable by fitting a linear equation to observed data. This model will lead to poor results because it will estimate probability less than zero and more than one which defies the principle of probability. We are exploring a classification problem; therefore, poor results were anticipated with this model. Support vector machine classifier uses a hyperplane to separate observations. Support vector machine classifier is ideal for linear data sets, however, in most real-world cases, data sets are non-linear.

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.

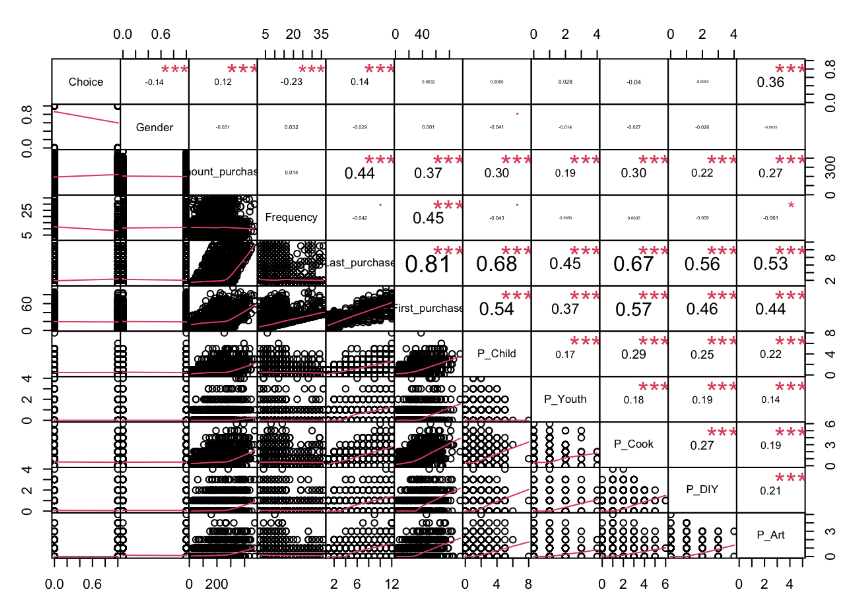
The support vector machine (SVM) is an extension of the support vector machine classifier that results from enlarging the feature space in a specific way, using kernels. Some assumptions are that the margins should be as big as possible and that support vectors are the most useful data points because they are the ones most likely to be classified incorrectly. Enlarging the feature space addresses non-linearity of the data. A kernel is a function that quantifies the similarity of two observations. There are various kernels used for SVM, which are linear, polynomial, sigmoid and radial. The advantage of using a kernel is computational, the data can be transformed into desired form with little computing cost. RBF, a radial kernel, was used for the case study, RBF is a non-linear kernel. For the radial kernel, only nearby training observations have an effect on the class label of a test observation. RBF is a popular kernel because it is localized (uses nearby training observations) and has a finite response along the complete x-axis. Three SVM models were used for the case study, linear, radial, and polynomial. The models were tuned for optimal hyperparameters, tune.svm function was utilized, gamma sequence was 0.01 - .5 by .01, cost sequence was .1 - 5 by .1. The tune.svm function tool took over an hour to generate, results set optimal cost at 2.1 and gamma was set at 0.03. Some of the limitations of the SVM model are because the classifiers put data points above and below classifying hyperplane, there is not a probabilistic explanation for classification; SVM will underperform if the number of features for each data point is more than the number of data samples in the training data; there can be a long training time for large datasets.

The Data

For this case analysis, we used a subset of the database available to BBBC. This consists of data for 400 customers who purchased the book and 1200 customers who did not, thereby posing an overrepresentation of the response group. This indicates a limitation of the dataset due to class imbalance. The class distribution is not equal and is skewed into one class. The data cleaning process revealed no missing values. The dependent variable for the analysis is *choice,* whether a book purchase was made or not. Below is a summary of the data prior to removing *observation*.

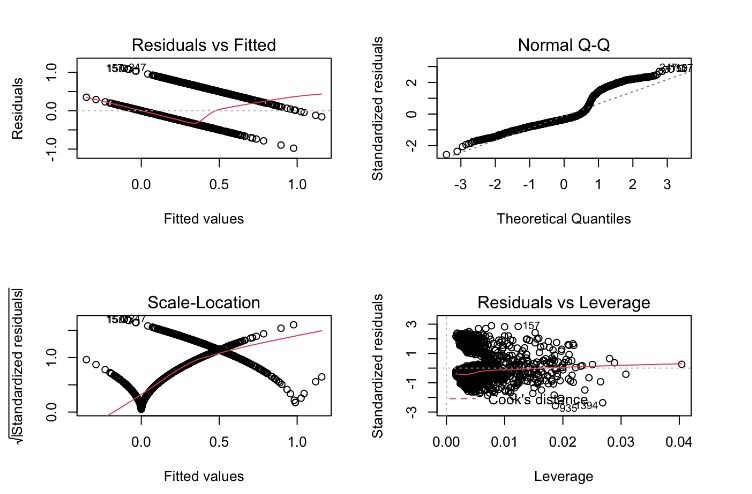


After inspecting our data, we determined that the *observation* variable is not significant to our models because it served as a placeholder, thus the variable was removed. Both c*hoice* and *gender* variables were binary and categorical, so they were set to factor variables. Below are distribution plots of customer orders where there appears to be a strong relationship in time between first and last purchase with a positive correlation of 0.81. There are no variables with correlations above 0.9 so this indicates a potential multicollinearity issue.



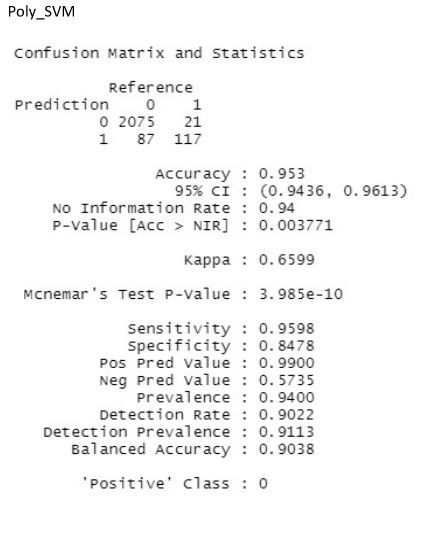
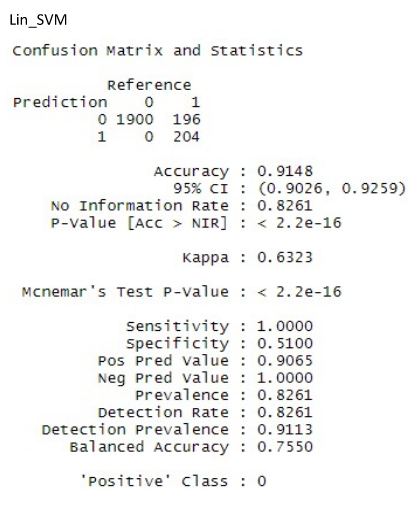
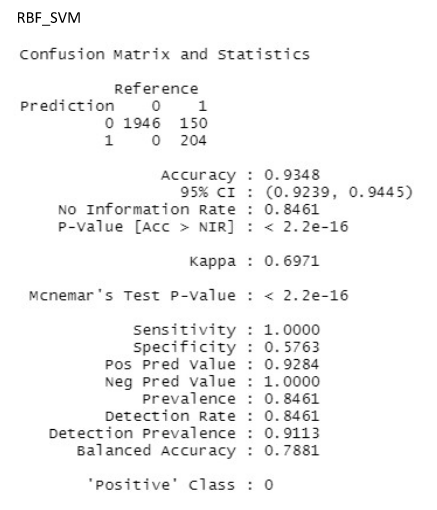
Findings

The mean square error between the final MLR model and the test data is exceedingly small, but it does not mean this is a good model. The disadvantage is the response variable, Choice is a classification variable that contains only 0’s and 1’s, thus, squaring it will not change anything. This also indicates the MLR model is not suited well for this classification dataset. Our model is non-linear. Linear regression’s predicted value is continuous, not probabilistic. Another disadvantage is that the linear regression model is sensitive to imbalanced data when using this method for classification problems. The plots below help illustrate why the MLR model is a poor fit. Take the Normal Q-Q plot for example. Instead of a straight line it is sigmoid shape. The Scale-Location plot makes trends in residuals more evident and can see a U-Shape trend in residuals. The Leverage plot graphs the standardized residuals against their leverage. This plot also includes Cook’s distance boundaries. Any point outside these boundaries is an outlier in the x direction. Since we cannot see the boundaries on our plot, we can conclude that there are no outliers.



The three SVM models used for the case study returned accuracy rates greater than 90%. The polynomial SVM model had the highest accuracy rate at 95.3%, followed by the radial model at 93.48% and then the linear SVM model at 91.48%. The response rate, customers that purchased the book compared to total 20,000 customers of the East campaign, was 9.03%. The test data set produced a response rate of 8.87%. Considering a mailing campaign for the Midwest of 50,000 customers, the estimated customers that would purchase the book would be 4,515 utilizing a response rate of 9.03%.

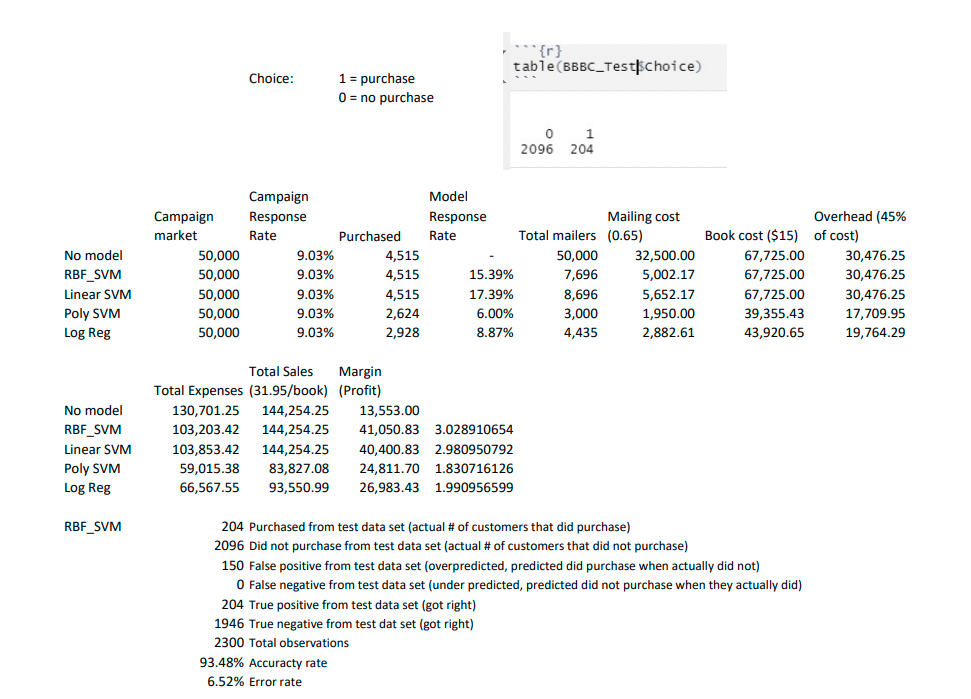
The radial and linear SVM models overfit the data, resulting in a response rate of 15.39% for the radial model and 17.39% for the linear model. The radial and linear SVM models predicted with 100% accuracy the true positives in the test dataset and had 0 false negatives. Selecting either radial or linear SVM for the mailing campaign would result in identifying all customers that would be expected to purchase the book and additional customers that would not purchase the book, false positives. In other words, if Bookbinder used either radial or linear SVM models for the Midwest campaign, the company would identify all customers expected to purchase the book, but also send additional mailings to customers that would not purchase the book. Conducting a naïve campaign, Bookbinder would mail to 50,000 potential customers, with radial the market would be 7,696 customers and with linear SVM it would be 8,696. Each campaign would capture all customers expected to buy the book 4,515, however with the radial or linear SVM Bookbinder would save in mailing costs.



For the polynomial SVM, the model underfit the data, it did not predict all the customers that would purchase the book and the model predicted that some customers would not purchase the book even though they did, false negatives. The response rate for the polynomial SVM model was 6.00%, however that rate is less than the expected response rate of 9.03%, indicating lost sales. If Bookbinder chose the polynomial SVM model, the market would be 3,000 customers resulting in 2,624 customers purchasing the book. The polynomial SVM model results in less sales and is not recommended as the model for the Midwest campaign.

The Logistical Regression model underfits the data, like the polynomial SVM model. The response rate for the Logistical Regression model was 8.87%, however that rate is less than the expected response rate of 9.03%, indicating lost sales. If Bookbinder chose the Logistical Regression model, the market would be 4,435 customers resulting in 2,928 customers purchasing the book. The polynomial SVM model results in fewer sales compared to the RBF and Linear SVM models and is not recommended as the model for the Midwest campaign.

Assuming a purchase response rate for the Midwest Campaign of 9.03%, 4,515 customers would purchase the book for a market campaign of 50,000 potential customers. Calculating total expenses based on the mailing, book and overhead costs that were given, Bookbinder would realize a profit of $13,553 if the company chose a naïve mailing campaign. We do not recommend Bookbinder pursue a naïve mailing campaign strategy; we suggest Bookbinder consider machine learning models as part of the Midwest Campaign marketing strategy. We estimate Bookbinder profit to range from $24K to $41K by pursuing different machine learning models. The RBF SVM model estimated a profit of $41,050, followed by Linear SVM with profit of $40,400. The Poly SVM and Logistical Regression models estimate profits at $24,811 and $26,983, respectively. Incorporating machine learning models in the Midwest Campaign would potentially result in 200% - 300% profit increase compared to a naïve campaign. The increase in profit is related to a reduction in mailing costs for the campaign, Bookbinder would not need to incur mailing costs for 50,000 potential customers. The models reduce mailing costs by predicting which potential customers Bookbinder should direct the Midwest Campaign to, from 50,000 potential customers to a range of 3,000 to 8,700 potential customers depending on which model was utilized.



Conclusion

In our analysis, there were three different modeling techniques utilized: linear regression, support vector machines (radial, linear, and polynomial kernels), and logistic regression. Out of all methods, SVM performed the best at predicting customers who did not purchase a book but lacked in predicting customers who did. This may be due to class imbalance, a limitation of the training dataset. SVM’s are incredibly sensitive to class distribution and choice of parameters. The logistic regression model did well in accurately predicting customers who purchased a book. Recommendation for future analyses would be to resample the data so that the class distribution is not so imbalanced. Another suggestion would be to use Random Forest, which is a type of SVM method that is less sensitive to class imbalance.

Reference

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.

Appendix

R code attached separately.