Bookbinder Study Case

Alex Martinez, Josh Gardner, Cameron Playle, and Guillermo Gallardo

2024-09-25

Paper Starts Here

Executive Summary

Brief introduction of problem. Summarizes key findings. Summarizes insights behind key findings.

Our Problem

Clear description of the problem, from an application and theoretical point of view. Outlines the report.

For this study case our goal is to evaluate how effective three different models are and compare them with the option of creating this campaign without a model. We are trying to determine which model will provide the best balance between cost savings and profit. By analyzing the performance of each model and comparing ti against the campaign, we will identify the most cost-effective approach that maximizes profit.

The three models we are comparing are the Linear Model (LM), Generalized Linear Model (GLM), and Support Vector Machine (SVM). Although we anticipate that the linear model may not perform well, we are still interested in understanding why it may not be the best fit for this case study. This exploration will help us gain valuable insights into the limitations of the linear model in this context and guide our decision-making process.

He said that we don't have to use the 50k population for our calculation to find which way would be best for the campaign. Email everyone in the 2300 or just a group of people based on our model.

Literature Review

Discusses and cites existing works in the theoretical and application realm.

Methods

Discusses types of variables, sample size, and sampling techniques (if any). Discusses the model(s) and its assumptions and limitations.

Our dataset was given to us divided into training and test sets. The training set includes 1,600 observations, while the test set contains 2,300 observations. The dataset consists of 12 variables, with one variable (observation) being removed. Two variables were converted into factors: gender and choice, with choice serving as our response variable. The remaining variables are numerical. For some of these numerical

variables, including Last_purchased, we will transform them into categorical variables to see how it would impact our modeling process.

Add stuff about unabalanced dataset?

GLM:

SVM: balance vs unbalanced comparison?

LDA maybe?

Data

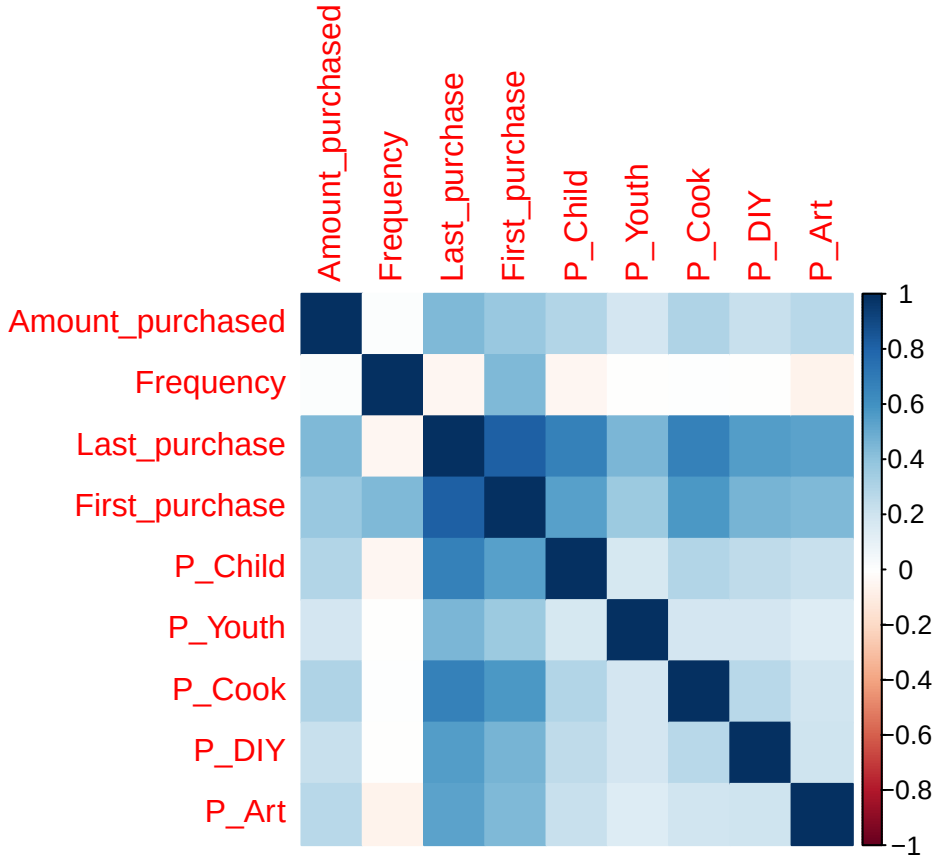
Discusses how data was handled, i.e. cleaned and preprocessed. Discusses distributions, correlations, etc.

Clean up

We modified two variables into factors, and we created categories for ADD STUFF HERE?

Correlation

The graph below highlights the variables with the highest correlations. We observed that *first_purchased* and *last_purchased* exhibit the strongest correlation. Based on this, we decided to create labels for these two variables to test their potential impact on our model's performance.. **ADD STUFF ABOUT RESULTS. DID IT WORK OR DID IT MAKE IT WORST?**



Results

Presents and discusses the results from model(s). Discusses relationships between covariates and response, if possible, and provides deep insights behind relationships in the context of the application.

GLM Results

Here we will be discussing the outputs from our logistic model we applied to the data set to best predict if someone will purchase a book - Choice (1/0). We first begin with all the variables in the data set to see which independent variables are significant.

```
##
## Call:
## glm(formula = Choice ~ . - Observation, data = bbc_train)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.3642284  0.0307411  11.848  < 2e-16 ***
## Gender           -0.1309205  0.0200303  -6.536 8.48e-11 ***
## Amount_purchased  0.0002736  0.0001110   2.464   0.0138 *
## Frequency        -0.0090868  0.0021791  -4.170 3.21e-05 ***
## Last_purchase     0.0970286  0.0135589   7.156 1.26e-12 ***
## First_purchase   -0.0020024  0.0018160  -1.103   0.2704
## P_Child          -0.1262584  0.0164011  -7.698 2.41e-14 ***
## P_Youth          -0.0963563  0.0201097  -4.792 1.81e-06 ***
## P_Cook           -0.1414907  0.0166064  -8.520  < 2e-16 ***
## P_DIY            -0.1352313  0.0197873  -6.834 1.17e-11 ***
## P_Art             0.1178494  0.0194427   6.061 1.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1434751)
##
##     Null deviance: 300.00  on 1599  degrees of freedom
## Residual deviance: 227.98  on 1589  degrees of freedom
## AIC: 1447
##
## Number of Fisher Scoring iterations: 2
```

With all the dependent variables, last_Purchase has the largest VIF so we decide to remove that from our model.

```
##
## Call:
## glm(formula = Choice ~ . - Observation - Last_purchase, data = bbc_train)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.3926595  0.0309609  12.682  < 2e-16 ***
## Gender           -0.1290720  0.0203424  -6.345 2.89e-10 ***
## Amount_purchased  0.0003518  0.0001122   3.135 0.001753 **
## Frequency        -0.0157943  0.0019980  -7.905 4.97e-15 ***
```

```
## First_purchase    0.0046036  0.0015884   2.898 0.003803 **
## P_Child          -0.0502183  0.0126891  -3.958 7.90e-05 ***
## P_Youth          -0.0225339  0.0175326  -1.285 0.198888
## P_Cook           -0.0667467  0.0131127  -5.090 4.00e-07 ***
## P_DIY            -0.0606486  0.0170835  -3.550 0.000396 ***
## P_Art             0.1916012  0.0167447  11.443  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1480058)
##
##     Null deviance: 300.00  on 1599  degrees of freedom
## Residual deviance: 235.33  on 1590  degrees of freedom
## AIC: 1495.8
##
## Number of Fisher Scoring iterations: 2

##           Gender Amount_purchased        Frequency   First_purchase
##         1.005634         1.235982         2.651820         7.182666
##          P_Child          P_Youth           P_Cook            P_DIY
##         1.949849         1.307915         2.009609         1.457362
##            P_Art
##         1.634878
```

First_purchase also has a large VIF so we will remove that from our model.

```
##
## Call:
## glm(formula = Choice ~ . - Observation - Last_purchase - First_purchase,
##     data = bbc_train)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.3731865  0.0302933  12.319  < 2e-16 ***
## Gender           -0.1263728  0.0203683  -6.204 6.99e-10 ***
## Amount_purchased  0.0003688  0.0001123   3.283  0.00105 **
## Frequency        -0.0112345  0.0012344  -9.101  < 2e-16 ***
## P_Child          -0.0275983  0.0100284  -2.752  0.00599 **
## P_Youth          -0.0014841  0.0159946  -0.093  0.92609
## P_Cook           -0.0428346  0.0102155  -4.193 2.90e-05 ***
## P_DIY            -0.0384262  0.0153017  -2.511  0.01213 *
## P_Art             0.2183323  0.0140081  15.586  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1486942)
##
##     Null deviance: 300.00  on 1599  degrees of freedom
## Residual deviance: 236.57  on 1591  degrees of freedom
## AIC: 1502.2
##
## Number of Fisher Scoring iterations: 2

##           Gender Amount_purchased        Frequency          P_Child
```

```
##         1.003526         1.232595         1.007587         1.212223
##          P_Youth           P_Cook            P_DIY            P_Art
##         1.083475         1.214043         1.163794         1.138879
```

Finally, P_Youth has a p-value > 0.05 so we will remove that to have our final model

```
##
## Call:
## glm(formula = Choice ~ . - Observation - Last_purchase - First_purchase -
##     P_Youth, data = bbc_train)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.3730697  0.0302577  12.330  < 2e-16 ***
## Gender            -0.1263681  0.0203619  -6.206 6.91e-10 ***
## Amount_purchased   0.0003679  0.0001118   3.289  0.00103 **
## Frequency         -0.0112341  0.0012341  -9.103  < 2e-16 ***
## P_Child           -0.0276675  0.0099975  -2.767  0.00572 **
## P_Cook            -0.0429132  0.0101772  -4.217 2.62e-05 ***
## P_DIY             -0.0385786  0.0152086  -2.537  0.01129 *
## P_Art              0.2182592  0.0139816  15.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1486016)
##
##     Null deviance: 300.00  on 1599  degrees of freedom
## Residual deviance: 236.57  on 1592  degrees of freedom
## AIC: 1500.2
##
## Number of Fisher Scoring iterations: 2
```

```
##           Gender Amount_purchased        Frequency          P_Child
##         1.003520         1.222450         1.007578         1.205526
##           P_Cook            P_DIY            P_Art
##         1.205706         1.150392         1.135282
```

Before we get into the confusion matrix. Lets explain the relationship of each independent variable to the dependent variable. We first compute the exponential of our coefficient ratio to get the odds ratio.

Table 1: Odds Ratios from the Logistic Model

	Variable	Odds Ratio
Gender	Gender (Male)	0.8812904
Amount_purchased	Amount Purchased	1.0003680
Frequency	Purchase Frequency	0.9888287
P_Child	Child Books Purchased	0.9727118
P_Cook	Cook Books Purchased	0.9579945
P_DIY	DIY Books Purchased	0.9621561
P_Art	Art Books Purchased	1.2439095

We observe the following key findings from the model:

- Gender (Male): Males decrease the odds of a client buying a book by a factor of 0.88.
- Amount of Books Purchased: A larger amount of books purchased increases the odds of a client buying a book by a factor of 1.
- Purchase Frequency: A higher frequency of books purchased decreases the odds of a client buying a book by a factor of 0.99.
- Child Books Purchased: A higher purchase of child books increases the odds of a client buying a book by a factor of 0.97.
- DIY Books Purchased: A higher purchase of DIY books increases the odds of a client buying a book by a factor of 0.96.
- Art Books Purchased: A higher purchase of art books increases the odds of a client buying a book by a factor of 1.24.

With this model, we then use it on our bbc_test sample to see how well it predicts and determine the most optimal cutoff to have the highest Sensitivity. Here are the following results:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2005   91
##          1  140   64
##
##                Accuracy : 0.8996
##                  95% CI : (0.8866, 0.9116)
##     No Information Rate : 0.9326
##     P-Value [Acc > NIR] : 1.000000
##
##                   Kappa : 0.3032
##
##  Mcnemar's Test P-Value : 0.001588
##
##             Sensitivity : 0.41290
##             Specificity : 0.93473
##          Pos Pred Value : 0.31373
##          Neg Pred Value : 0.95658
##              Prevalence : 0.06739
##          Detection Rate : 0.02783
##    Detection Prevalence : 0.08870
##       Balanced Accuracy : 0.67382
##
##        'Positive' Class : 1
##


## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2095    1
##          1  199    5
##
```

```
##                Accuracy : 0.913
##                  95% CI : (0.9008, 0.9242)
##     No Information Rate : 0.9974
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0428
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.833333
##             Specificity : 0.913252
##          Pos Pred Value : 0.024510
##          Neg Pred Value : 0.999523
##              Prevalence : 0.002609
##          Detection Rate : 0.002174
##    Detection Prevalence : 0.088696
##       Balanced Accuracy : 0.873293
##
##        'Positive' Class : 1
##
```

We see the best cutoff for the highest sensitivity is at 0.8. With this, the performance of our model on the bbc_test data set is 91% accurate overall, and our sensitivity is 83% with a specificity of 91%

Conclusion

Concludes with a summary of the aim and results. Discusses alternative methods that can be used.

Based on our analysis, the **ADD MODEL**

Add details on if it is better to use model or just send campaign to everyone in the list.

STUFF FROM THE PDF

Summarize the results of your analysis for the three models. The training, testing, and prediction data can be found on Blackboard.

Interpret the results of the models. In particular, for models the influential covariates and their coefficients, provide insights.

BBBC is considering a similar mail campaign in the Midwest where it has data for 50,000 customers. Such mailings typically promote several books. The allocated cost of the mailing is $0.65/addressee (including postage) for the art book, and the book costs $15 to purchase and mail. The company allocates overhead to each book at 45% of cost. The selling price of the book is $31.95. Based on the model, which customers should Bookbinders target? How much more profit would you expect the company to generate using these models as compare to sending the mail offer to the entire list.

Please also summarize the advantages and disadvantages of the three models, as you experienced in the modeling exercise. Should the company develop expertise in either (or all) of these methods to develop in-house capability to evaluate its direct mail campaigns.

How would you simplify and automate your recommended method(s) for future modeling efforts at the company?