EXECUTIVE SUMMARY: DIRECT MARKETING IN THE AGE OF BIG DATA

PURPOSE OF PROJECT

Through this project, we hope to provide data-driven and strategic targeting recommendations to Antixo, a direct marketing firm that currently mails catalogs to its entire customer base of 5 million individuals. We have two objectives to consider: target as many actual buyers as possible and as little non-buyers as possible, and target the proportion of customers that maximize the firm's profit. For the first objective, we are assuming the firm will only target 30% of its customers.

OUR RECOMMENDATIONS

To satisfy our first objective--accurately targeting buyers and avoiding non-buyers--we suggest the firm induce a Bagging model and use the results to classify customers by their probability of purchase. We suggest this model because our analysis showed that this particular model--when compared to K-Nearest Neighbor, J-48, Random Forest, and NaiveBayes models--yielded the highest classification accuracy. For the second objective--targeting a subset of customers that maximizes the firm's profitability--we suggest the firm induce a NaiveBayes model and target customers above the score threshold. We chose this model because when profits and costs for each model were plotted against each other, NaiveBayes yielded the highest potential profit.

The main difference between the two modeling techniques we used was that the first objective only relied on two variables to determine whether or not a customer would make a purchase, while the second objective considered all variables significant in maximizing profit. This can be explained by the difference between the two objectives--as one was geared towards accuracy and the other towards profitability. It was for this very reason that we also choose two different methods to evaluate the success of our models. For the first objective, since all errors were considered equal, we used classification accuracy rate as an evaluator. In contrast, for the second objective we used profit charts and a cost/benefit analysis to pick the model that yielded the highest profit. We did so because we recognized that different errors had different costs. Thus, although a model may have the highest classification accuracy rate it may not necessarily maximize profit.

JUSTIFICATION

Antixo has done well in collecting vast amounts of data about its customers, and is currently missing out on a significant opportunity to leverage it. It is in the company's best interest--both financially and for efficiency's sake--to learn from its historical data and build predictive models. Both models, although not perfect solutions, prove to produce better results than targeting its entire customer base. With printing and mailing costs increasing, it is more important than ever for Antixo to understand how to make its money work efficiently and by inducing the models we've built and following our recommendations, we believe the firm can do so.

DIRECT MARKETING IN THE AGE OF BIG DATA

PROBLEM

We are working with Antixo, a direct marketing firm that mails catalogs to its customer base of 5 million individuals. In light of considerable increases to its direct mailing and printing costs, the firm has decided to move away from its current mass-mailing strategy. Rather than target specific customers based on their estimated likelihood of response or money spent, this method simply mails catalogs to all 5 million customers. Antixo would like to replace its existing strategy with one that will target a subset of customer, helping the firm meet two separate objectives for its mailing campaigns.

The first objective is to reach out to as many actual buyers as possible and to as little non-buyers as possible. Assuming the firm will only target 30% of its customers in the future, Antixo would like to maximize the accuracy at which it classifies them. Separately from accuracy, it would also like to target a subset of customers that will maximize profitability. We used the data provided to us to build models that predict whether a customer will make a purchase or not and how much money they will spend on their orders, so that Antixo can make data-driven decisions that will ultimately reduce costs and increase profits.

DATA PREPARATION

Our data consisted of customer information from Antixo's most recent mass-mailing campaign, including customers' responses (or lack thereof), the margin (prior to subtracting the mailing costs) from each order. and other attributes relating to the customer's purchases such as the dollar amount of their most recent order, the date of their most recent order, and their average amount spent. A complete list of all of the attributes and the target variables can be seen in *Figure 1*. We used WEKA to prepare all of our data and to induce our predictive models. Since Target Variable B--whether the customer made a purchase or not--and Target Variable D--the customer's order margin in the last campaign--are dependent on one another, we decided to split the data into two sets. One set contained all the attributes with Target Variable B and the other set contained all the attributes with Target Variable D. The dataset with Target Variable B was used to induce models for the first objective, while the dataset with Target Variable D was intended to be used to create a model to satisfy the second objective.

The next step we took in preparing our data was testing the information gain of the attributes in regards to their specified target variables. We wanted to measure the worth of the attributes in respect to their classes to see if we needed to remove any insignificant attributes before creating our models. The information gain evaluator is a way to get a ranked list of the most predictive features according to their information gain scores. A score of 0 indicates that the attribute is not significant in regards to the class. As seen in *Figure 2,* when we evaluated all the attributes (1-8 in *Figure 1*) against Target Variable B, every attribute led to an information gain, but some results were higher than others. The information gain attribute selector was beneficial in this case because Target Variable B is a binary class indicating whether the customer made a purchase or not. In contrast, Target Variable D is a numeric class so we could not perform the information gain evaluation.

Finally, to avoid an over-optimistic evaluation, we randomly partitioned each of our two data sets into training and test sets. We divided each data set into three folds. One fold was held out when building the model and later used as a test set, while the other two folds were used to induce the model.

DATA ANALYSIS

Objective 1

As stated in the problem section, the first objective the company wants to achieve is to reach out to as many buyers as possible while also avoiding non-buyers. By using historical data to predict whether or not a customer will make a purchase, we are able to selectively target the top 30% of customers with the highest probability of making a purchase. We used the Target Variable B dataset to create predictive models, and relied on the classification accuracy rate as an evaluator as to which model would serve the company best. The classification accuracy rate gives us the proportion of examples whose class is predicted accurately by the model. To determine whether or not our model was better than a trivial classification rule, we compared the classification accuracy rate to the base rate. In doing so we found that of the 5,131 customers we collected data from, precisely 3,083 customers did not make a purchase from the last catalog--resulting in a base rate of 60.09%. This means that at random, we could predict that 60% of customers wouldn't respond, so our classification accuracy rate had to be higher than this to be considered a good model.

Once we determined the base rate, we used all of the attributes in our Target Variable B data set to create J48, Bagging, Random Forest, Naive Bayes, and K-Nearest Neighbor models. In the test options section, we made sure to supply the test data that we had previously prepared. Our results were disappointing, as the highest classification accuracy rate returned from all five models was 59.61%, which was lower than the base rate. After looking over our information gain results and running multiple iterations of deleted variables, re-splitting the data into training and test sets, and then running the models again, we finally found the two-variable combination that led to the highest classification accuracy rates: amount of most recent order and order amount category for the last 24 months. The summary output of each model can be seen in *Figures 3-7*.

The model that produced the highest classification accuracy rate was the Bagging model, with 62.0105% (See *Figure 3*). Although this is only slightly higher than the base rate, we can conclude that this is a good model because it is more accurate than randomly assigning a customer to "would not respond" with a 60% probability. The Random Forest model (*Figure 7*) and K-Nearest Neighbor model (*Figure* 6) were also close, as their classification accuracies were only lower by 0.2922%, and 0.3507% respectively. Therefore, we turned to precision and recall evaluation methods to choose the best model. Although Random Forest had a slightly higher recall than Bagging, Antixo's main focus for this objective was how many of the predicted buyers actually bought something from the catalog, or in other words the model's precision. Since Bagging had the highest precision and classification accuracy rate, we chose it as our final model.

Objective 2

In order to maximize profitability, we needed to create a model that predicts the customers who are going to spend the most on the catalog, and then mail the catalog to the top 30% of customers with the highest probabilities of making those purchases. For this model, we originally planned on using the dataset with Target Variable D--the order margin from the last campaign. We split this data into training and test sets during our data preparation.

Since classification accuracy rate implies that all errors are equally as bad, this would not be a sufficient model evaluation method. Instead, we used cost-sensitive analysis through profit and lift charts to choose the model that yielded the highest profitability for Antixo. The cost of sending the catalog to a person who does not make a purchase can be quantified as -$4.00, since the mailing and printing costs used to produce the catalog is $4.00. For example, if they send the catalog to a person who does not end up buying anything, they have lost $4.00 from production and mailing costs, and gained $0.00. In terms of the benefit from correctly targeting a buyer, we decided to choose $9.00 dollars. We used this number by taking the average of Attribute 6, average amount spent by customer, which ended up being $12.76. We then had to factor in the $4.00 mailing costs, which brought us to an average profit of about $9.00 dollars per customer.

Using this information, we ran a Linear Regression, K-Nearest Neighbor, Bagging, and Random Forest models with all attributes on Target Variable D. After doing so, we realized that we would not be able to run a cost-benefit analysis on the data because our target was numeric instead of binary. We also found that the correlation of the models produced was very low. It then occurred to us that we could still use Target Variable B data set to predict which customers were most likely to respond, and account for the profits and losses by using the cost-benefit analysis. Similar to our process for Objective 1, we ran multiple models varying the number and selection of variables. This time though, we compared the maximum profits from each model using the cost/benefit profit chart instead of comparing classification accuracy rates. The model that yielded the highest profit was NaiveBayes (*Figure 10*). The model used Target Variable B as its class, and was induced with all attributes 1-8 (*Figure 1*) to predict the class. We tried using less attributes to see if this would maximize profitability, but the profits turned out to be lower. We created a graph (*Figure 13*) that contains the five final profit charts we based our recommendation on. As one can see from the graph, K-Nearest Neighbor and J-48 produced the lowest profitability and when compared to targeting the population at random, they produced no gain (see *Figures 9 and 11)*.

RECOMMENDATIONS

Objective 1

In order to effectively target a subset of customers who are most likely to buy something from the catalog, we need to figure out the minimum probability threshold at which our model correctly predicts the customer will make a purchase. To do this, we output the Bagging Model's individual predictions for each customer in WEKA. Then, we transferred this into Excel where we sorted the customers from the highest probability of buying something to to the lowest probability. For this particular instance, we found that Antixo would be fairly successful targeting a customer who had a 0.457 probability or higher of buying. If it wanted to be almost 100% accurate, the firm would have to raise the probability threshold to 0.611 or higher.

Since Antixo already has a target in mind--a subset of only 30% of their customers--this new strategy will relieve them of the cost of sending catalogs to all 5 million of their customers. When it comes time to mail out the catalogs, Antixo should simply calculate the probability that each customer will buy something from the catalog using the Bagging Model we created, and then rank every customer according to their probability of purchase from largest to smallest in Excel. The top 30% of customers on that list have a higher probability of making a purchase than anyone else who were to receive a catalog. This, in turn, would satisfy its first objective of targeting as many buyers as possible and would avoid targeting non-buyers. If the firm did not have the 30% subset in mind and instead wanted to find the probability at which all customers who were targeted would say yes, it would have the option to do so as well (using the same method mentioned above). Instead of targeting the top 30%, it would pick the minimum probability threshold at which all customers who were predicted to buy something bought something and not target anyone who had a probability lower than that number.

Objective 2

In the interest of targeting those who yield the highest probability, the firm should induce our NaiveBayes model and target those customers with a probability of buying at a threshold of 0.1575 or higher. In doing so, our model predicts that Antixo will obtain the highest possible profit. Understanding that costs and benefits may change depending on a number of factors outside of the scope of this project, we suggest that it constantly monitor and update the cost and benefit portion of this model to ensure its predictions are accurate. Factors that may change include printing costs, mailing costs, and the benefit received from customers who are targeted as buyers and actually buy from the catalog. By putting our model to use, Antixo can increase their profit by $353.96 when compared to targeting the population at random, achieving a maximum profit of $2,138. The probability threshold at which the company chooses which customers to target may change as well, depending on the costs used in the cost-benefit analysis, so the firm should make sure to keep an eye on this if conditions change.

APPENDIX

Figure 1. Target Variables and Attributes

1. RAMNTALL: Total order amounts in dollars in the last 3 years
2. NBUYALL: number of orders over the past 3 years
3. LASBUY: Amount of most recent order
4. LASDATE: Date of most recent order (Year and month)
5. FIRSTDATE: Date of first order (Year and month)
6. AVGBUY: Average amount spent by customer
7. RFA2_F: Frequency of order over the last 24 months
   1, 2, 3, or (4 or more)

8. RFA2_A: Order amount category for the last 24 months
   A=$0.01 - $1.99
   B=$2.00 - $2.99
   C=$3.00 - $4.99
   D=$5.00 - $9.99
   E=$10.00 - $14.99
   F=$15.00 - $24.99
   G=$25.00 and above

9. Target B: Order in last campaign, i.e., customer's buying decision (1: buy, 0: not buy)
10. Target D: Order margin in the last campaign. Note that only the item's cost to the firm has been considered to compute this margin, and the mailing cost has not been subtracted from the revenue to compute the margin. Also note the order amount is always zero when no purchase was made.

Figure 2. Infogain for Target Variable B

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 TARGET_B):
        Information Gain Ranking Filter

Ranked attributes:
 0.01825   8 RFA_2A
 0.01711   3 LASTBUY
 0.01681   6 AVGBUY
 0.01469   7 RFA_2F
 0.01064   2 NBUYTALL
 0.00733   5 FISTDATE
 0.00465   4 LASTDATE
 0.00387   1 RAMNTALL

Selected attributes: 8,3,6,7,2,5,4,1 : 8
```

Figure 3. Bagging (30 Iterations, J48)

```
=== Summary ===

Correctly Classified Instances        1061               62.0105 %
Incorrectly Classified Instances       650               37.9895 %
Kappa statistic                          0.1071
Mean absolute error                      0.4663
Root mean squared error                  0.4833
Relative absolute error                 97.215  %
Root relative squared error             98.6934 %
Total Number of Instances             1711

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.908    0.813    0.627      0.908    0.742      0.138    0.587     0.653     0
                 0.187    0.092    0.574      0.187    0.283      0.138    0.587     0.470     1
```

Figure 4. NaiveBayes

```
=== Summary ===

Correctly Classified Instances        1010               59.0298 %
Incorrectly Classified Instances       701               40.9702 %
Kappa statistic                          0.1268
Mean absolute error                      0.4644
Root mean squared error                  0.4874
Relative absolute error                 96.8141 %
Root relative squared error             99.5309 %
Total Number of Instances             1711

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.702    0.578    0.646      0.702    0.673      0.128    0.594     0.663     0
                 0.422    0.298    0.485      0.422    0.451      0.128    0.594     0.477     1
```

Figure 5. J48

```
=== Summary ===

Correctly Classified Instances        1052               61.4845 %
Incorrectly Classified Instances       659               38.5155 %
Kappa statistic                          0.0843
Mean absolute error                      0.4676
Root mean squared error                  0.484
Relative absolute error                 97.4774 %
Root relative squared error             98.8245 %
Total Number of Instances             1711

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.849    0.621      0.923    0.742      0.117    0.583     0.650     0
                 0.151    0.077    0.566      0.151    0.238      0.117    0.583     0.458     1
```

Figure 6. K-nearest neighbor (k=1)

```
=== Summary ===

Correctly Classified Instances        1055               61.6598 %
Incorrectly Classified Instances       656               38.3402 %
Kappa statistic                          0.1013
Mean absolute error                      0.4627
Root mean squared error                  0.4864
Relative absolute error                 96.462  %
Root relative squared error             99.3279 %
Total Number of Instances             1711

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.810    0.626      0.900    0.738      0.129    0.590     0.661     0
                 0.190    0.100    0.558      0.190    0.284      0.129    0.590     0.480     1
```

Figure 7. Random Forest (100 Iterations, NumFeature = 0)

```
=== Summary ===

Correctly Classified Instances        1056               61.7183 %
Incorrectly Classified Instances       655               38.2817 %
Kappa statistic                          0.1035
Mean absolute error                      0.463
Root mean squared error                  0.4857
Relative absolute error                 96.5199 %
Root relative squared error             99.1754 %
Total Number of Instances             1711

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.899    0.807    0.626      0.899    0.738      0.131    0.590     0.661     0
                 0.193    0.101    0.559      0.193    0.287      0.131    0.590     0.479     1
```

Figure 8. Bagging Cost-Benefit Chart

Figure 9. J-48 Cost-Benefit Chart

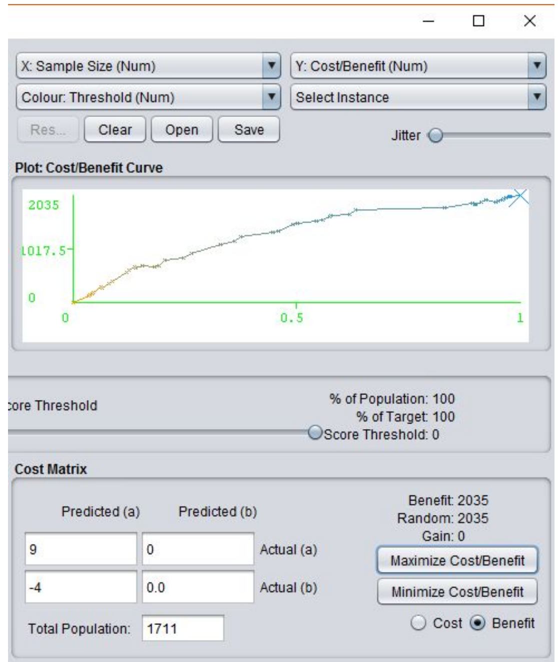

Figure 10. NaiveBayes Cost-Benefit Chart

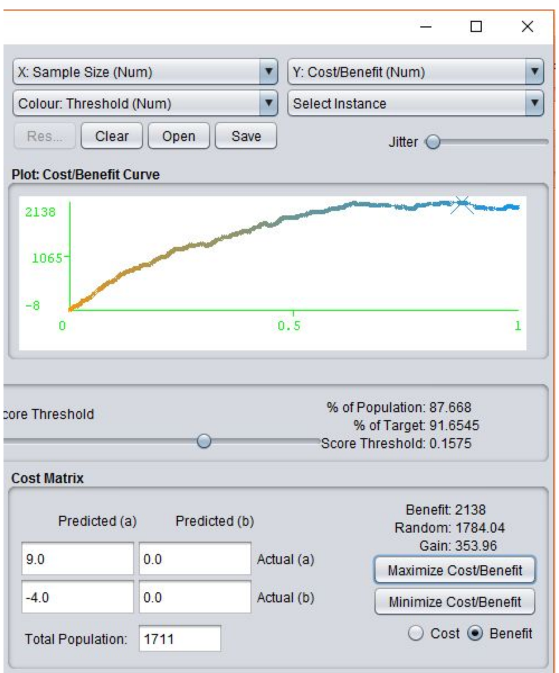

Figure 11. K-NN Cost-Benefit Chart

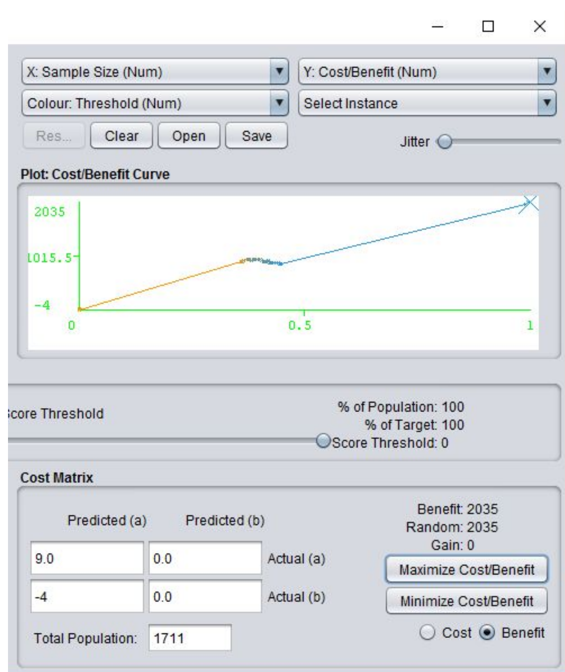

Figure 12. Random Forest Cost-Benefit Chart

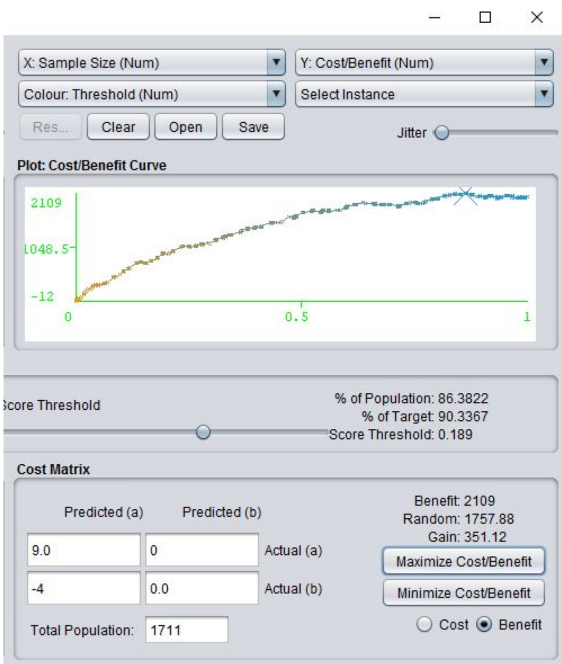

Figure 13. Comparison of Profit Lift Charts