***North-Point Software Listing Company*:**

***Customer Selection***

Lavanya Govindarajan

Department of Computer and Information Sciences, Webster University

May 08, 2024

Table of Contents

[1. INTRODUCTION 3](#_Toc166067096)

[**1.1** **BUSINESS GOAL** 4](#_Toc166067097)

[**1.2** **ANALYTICS GOAL** 4](#_Toc166067098)

[**1.3** **ANALYTICS APPROACH** 5](#_Toc166067099)

[2. EXPLORATORY DATA ANALYSIS 6](#_Toc166067100)

[**2.1** **ZERO VALUE CHECK** 9](#_Toc166067101)

[**2.2** **OUTLIER ANALYSIS** 10](#_Toc166067102)

[**2.3** **NUMERICAL VARIABLE ANALYSIS** 11](#_Toc166067103)

[3. PREDICTOR ANALYSIS AND RELEVANCY 18](#_Toc166067104)

[4. DIMENSION REDUCTION 23](#_Toc166067105)

[5. DATA PARTITIONING 24](#_Toc166067106)

[6. DATA NORMALIZATION 25](#_Toc166067107)

[7. CLASSIFICATION MODEL SELECTION 26](#_Toc166067108)

[8. REGRESSION MODEL SELECTION 30](#_Toc166067109)

[9. DATA SAMPLING 33](#_Toc166067110)

[10. PREDICTION 34](#_Toc166067111)

[11. CONCLUSION 36](#_Toc166067112)

[12. FUTURE WORK 37](#_Toc166067113)

1. **INTRODUCTION**

North-Point initially established itself as a software manufacturer specializing in games and educational software. As part of its evolution, it expanded its offerings to include third-party titles, aiming to provide a comprehensive solution for its customers' business software needs. This approach enabled North-Point to address a broader spectrum of customer requirements while concentrating its internal resources on areas where it possessed the greatest expertise and differentiation. Moreover, it strengthened its market position by offering a more extensive suite of software solutions compared to competitors who only offered proprietary software.

In a recent development, alongside its core offerings in games and educational software, North-Point transitioned into a software listing firm, with a primary focus on providing information and recommendations to users, aiding them in selecting appropriate software solutions tailored to their needs. With years of cultivating relationships with diverse customers, North-Point regards its customer base as a pivotal asset. To leverage and expand this asset further, a strategic decision was made to join a consortium comprising other specialized listing firms in software and hardware products. This consortium presents an advantageous avenue for acquiring additional trustworthy customers, as it comprises some of the most reputable listing firms, each bringing their esteemed customers to the table.

As part of this consortium arrangement, North-Point has contributed 200,000 customers to a shared pool, which currently encompasses a total of 5,000,000 customers. In return, North-Point is entitled to select 200,000 customers from this pool. To ensure fairness, the consortium permits its members to employ predictive analytics methodologies to make informed decisions regarding customer selection from the shared pool.

* 1. **BUSINESS GOAL**

In pursuit of refining its customer acquisition strategy, North-Point conducted a mailing experiment. About 20,000 customers were randomly selected from the pool and sent mails containing recent listings at $2 per recipient, incurring a total cost of $40,000. Encouragingly, the outcomes were favorable, with 1065 customers, constituting approximately 5.3% of the total recipients, making purchases. This translated to a total turnover of $205,249, resulting in a gross profit of $165,249. Despite the profitability of this experiment, it was still based on chance, much like gambling, where outcomes are unpredictable.

The business goal of North-Point is to identify top 180,000 profitable customers from the pool, thereby maximizing its profits. Additionally, North-Point aims to mitigate the influence of randomness in customer selection through the deployment of predictive algorithms.

* 1. **ANALYTICS GOAL**

The analytics goal is derived from the overarching business objective of North-Point, which aims to optimize resource allocation and forecast profits effectively. To achieve this, the analytics approach encompasses two key objectives:

**Customer Identification:** Utilize classification algorithms to assess the probabilities of a customer being a purchaser or non-purchaser. This enables North-Point to concentrate resources on potential buyers, maximizing the efficiency of marketing and sales efforts.

**Sales Prediction:** Employ regression analysis to forecast the amount each customer is likely to spend on purchasing products. This predictive model facilitates profit and cost forecasting, aiding in strategic planning and decision-making for each quarter.

The analytics process begins with data extraction from a pool of 5 million records, representing North-Point's customer base. To effectively prioritize resources, the classification algorithm is utilized to rank the top 180,000 customers who are potential purchasers. This approach optimizes resource allocation by focusing efforts on the most promising prospects. Following customer identification, regression models are deployed to predict sales for the 180,000 potential purchasers. By estimating individual spending amounts, North-Point gains insights into potential revenue streams and can tailor strategies accordingly. The performance of classification and regression models is assessed against a baseline model, which involves randomly selecting 180,000 customers. This comparison provides insight into the effectiveness of predictive analytics in improving customer targeting and sales forecasting.

* 1. **ANALYTICS APPROACH**

The approach involves 3 phases. In Phase I, analysis will be done on 2000 customers from the 20,000 extracted customers by creating a balanced dataset of 1000 purchasers and 1000 non-purchasers. The balanced dataset will help ensure unbiased results. A classification algorithm will be used to gauge the probabilities of a purchaser. A threshold will be set by the firm as per its requirement, and all the customers with probabilities above the threshold will be identified as purchasers. A regression algorithm will then be applied to the purchasers alone to forecast the amount of spending. This will then be compared with a naïve base model (numbers obtained through random selection) to determine the best model that outperforms the naïve base model.

As a precautionary step, in Phase II, the selected models will then be applied to the remaining 18,000 dataset, which will be unbalanced and represent real-world data for easy comparison on how well the model's prediction matches the real-world numbers.

In the third and final phase, Phase III, the classification model will be deployed on the pool of 5,000,000 to rank top 180,000 customers in the decreasing order of their probability of being a purchaser. Then the regression model will be applied to these 180,000 customers to estimate the total spending. Based on this spending estimation, the gross profits can be calculated, and the company can plan its quarterly spending, modify listings, enhance marketing efforts, etc., accordingly.

1. **EXPLORATORY DATA ANALYSIS**

The dataset comprises 20,000 customer records, including 1000 purchasers and 1000 non-purchasers. Creating such a balanced dataset allows us to train our classification algorithm more effectively, enabling it to learn patterns and characteristics associated with both types of customers equally. This helps prevent the model from becoming skewed towards the majority class and ensures accurate prediction of the likelihood of a customer being a purchaser or non-purchaser, irrespective of class distribution. There are two outcome variables in the dataset, *Purchase* and *Spending*. A comprehensive description of the 25 attributes is provided in Figure 2.1.

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Sequence\_number | A unique number generated for each customer in the dataset. |
| US | A binary variable with a 1 representing customers from the US and 0 representing customers from other countries. |
| Source\_a,source\_c, source\_b, source\_d, source\_e, source\_m, source\_o, source\_h, source\_r, source\_s, source\_t, source\_u, source\_p, source\_x, source\_w | These are  binary variables representing different sources from which customers were acquired, with 1 indicating the customer’s origin from a particular source. |
| Freq | A numerical variable recording the number of purchases made by the customer last year at the source. |
| Last\_update\_days\_ago | A numerical variable indicating the number of days since the last update with the customer's data. |
| 1st\_update\_days\_ago | A numerical variable indicating the number of days since the first update with the customer's data. |
| Web order | A binary variable with a 1 indicating a purchase made by the customer through a web order and 0 indicating other mode of purchase. |
| Gender=male | A binary variable with a 1 indicating a male customer and 0 indicating a female customer. |
| Address\_is\_res | A binary variable with a 1 indicating a residential address and 0 indicating a non-residential address. |
| Purchase | A binary variable with a 1 indicating a purchaser and 0 indicating a non-purchaser in test mail. |
| Spending | A numerical variable indicating the amount spent in dollars by customers who made a purchase in test mail. |

**FIGURE 2.1** Attribute Description.

It is important to note that the attributes *Purchase* and *Spending* may not exist in the original dataset for the 5,000,000 customers. These two attributes were added by North-Point based on the results of the mailing experiment. It is assumed that the remaining attributes are present in the original dataset. For simplicity and readability, the attributes are renamed by removing white spaces, special characters, and using camel casing for better readability. The mapping of the old and new columns is represented in Figure 2.2.

|  |  |
| --- | --- |
| **Old Attribute Names** | **New Attribute Names** |
| sequence\_number | sequenceNo |
| US | USCustomer |
| source\_a, source\_b, source\_c, source\_d, source\_e, source\_m, source\_o, source\_h, source\_r, source\_s, source\_t, source\_u, source\_p, source\_x, source\_w | sourceA, sourceB, sourceC, sourceD, sourceE, sourceM, sourceO, sourceH, sourceR, sourceS, sourceT, sourceU, sourceP, sourceX, sourceW |
| Freq | frequency |
| last\_update\_days\_ago | lastUpdate |
| 1st\_update\_days\_ago | firstUpdate |
| Web order | webOrder |
| Gender=male | gender |
| Address\_is\_res | isResAddr |
| Purchase | purchase |
| Spending | spending |

**FIGURE 2.2** Old and new attribute names mapping.

The data analysis was conducted using the R programming language, chosen for its widespread usage in analysis and its vast array of packages and applications. The skim function output, as shown in Figure 2.3, indicates that all attributes are numerical. No missing values or NAs were found in the data, indicating good data quality. Upon closer inspection, most attributes such as *USCustomer,* the *sources, webOrder, gender, isResAddr,* and *purchase* are binary in nature, taking values of either 1 or 0. Since they are already in binary format, they do not require explicit factorization. The attribute *sequenceNo* is numeric and continuous, serving as a unique identifier. However, it might not be particularly useful in modeling due to its lack of meaningful information or predictive power.

**A screenshot of a computer

Description automatically generated**

**FIGURE 2.3** Skim function output.

* 1. **ZERO VALUE CHECK**

The non-binary numerical attributes were then checked for zeros, revealing an anomaly. The customer with sequence\_number *711,* identified as a non-purchaser, was expected to have made no purchases. However, the Spending column indicated a purchase of $1 for this customer as shown in Figure 2.11. Consequently, the Spending column was rectified and set to 0 for this specific customer to maintain a balanced dataset of 1000 purchasers and 1000 non-purchasers.

A graph with numbers and a line

Description automatically generated

**FIGURE 2.11** Anomaly: Non-Purchaser with Spending = $1

* 1. **OUTLIER ANALYSIS**

Box plots were created for the remaining four numerical attributes to identify outliers. As depicted in Figure 2.21, the attribute *frequency* exhibits some outliers. However, it is important to retain these outliers as they provide valuable insights into the distribution and behavior of the data. Similarly, the attribute *spending* also displays outliers, particularly in the range of 1000 to 1500. Despite this, it is retained as these outliers may represent significant spending behavior that should not be overlooked.A row of different colored lines

Description automatically generated with medium confidence

**FIGURE 2.21** Box plot for numerical variables

* 1. **NUMERICAL VARIABLE ANALYSIS**

Summary statistics were computed for the non-binary variables. Descriptive statistics for frequency, as shown in Figure 2.31, indicate a range from 0 to 15, suggesting that some customers haven't made any purchases while others have made up to 15. The mean, approximately 1.417, and median, 1, reveal that most values are relatively low. Specifically, 25% of observations have a frequency of 1 or less, and 75% have a frequency of 2 or less. The histogram in Figure 2.31 illustrates a right-skewed distribution of frequency, with very few observations having a frequency greater than 5.

A screenshot of a graph

Description automatically generated

**FIGURE 2.31** Histogram and summary statistics for frequency.

From the summary statistics of the attribute *lastUpdate* in Figure 2.32, it can be inferred that the number of days since the last update on the customer's data ranges from 1 to 4188 days, with a median value of 2280. Additionally, the data distribution, as shown in the histogram in Figure 2.32, appears to be slightly positively skewed, as the mean (2155) is slightly lower than the median. Approximately 75% of the data fall under 3139 days since the last update. On average, the customer data hasn't been updated for approximately six years. This suggests that the firm's recent interaction with customers might be infrequent or outdated. A six-year period without interaction seems considerably lengthy and could indicate potential opportunities for re-engagement or updated outreach strategies.**A graph of a bar graph

Description automatically generated with medium confidence**

**FIGURE 2.32** Histogram and summary statistics for lastUpdate.

From the summary statistics of the attribute *firstUpdate* in Figure 2.33, it is evident that the variable spans from a minimum value of 1 to a maximum value of 4188 , representing the number of days since the first update, which is same as that of the attribute *lastUpdate*. The median value of 2721 indicates that half of the data points fall below this value, while the mean value of 2436 suggests the average number of days since the first update. This distribution suggests that the majority of updates occurred within a span of 1682 to 3353 days which is approximately between 4.5 to 9 years. The data reveals that the oldest first update on a customer is approximately 11.5 years ago, while the average first update is around 6.5 years ago. This suggests a significant longevity in the firm's relationships with its customers, with most customers having remained with the firm for an extended period. Such enduring customer relationships can indicate a high level of satisfaction, loyalty, and trust in the firm's products or services.

**A graph of a bar graph

Description automatically generated with medium confidence**

**FIGURE 2.33** Histogram and summary statistics for firstUpdate.

Extracting only the customers with "*purchase* = 1" indicates those who made purchases, as for the non-purchasers, the spending value would be zero. The summary statistics of the outcome variable *spending* in Figure 2.34, shows that the amount in dollars spent by the customers to make a purchase, spans from a minimum of $1 to a maximum of $1500. Approximately, 75% of the spending is below $233 and the average appending observed is $205. This summary reveals a concentration of spending towards lower amounts(<$500).

**A graph of a number of numbers

Description automatically generated with medium confidence**

**FIGURE 2.34** Histogram and summary statistics for spending (only for purchasers).

The distributions of the attributes USCustomer, webOrder, gender, and isResAddr are analyzed through bar charts to effectively represent the binary nature of these attributes. From the bar charts in Figure 2.35, it is inferred that approximately 82% of the customers are US-based, indicating a dominance of customers from the USA in the customer base. Additionally, about 42% of the customers have placed their orders through the web, with the majority opting for other methods. Although there isn't a significant disparity between male and female customers, the number of male customers is slightly higher than that of females. A group of bar graphs

Description automatically generated with medium confidence

**FIGURE 2.35** Bar Charts for USCustomer, webOrder, gender and isResAddr.

The majority of customers were acquired through a *sourceE* representing approximately 15% of the total customer base. Conversely, *SourceP* accounted for the smallest portion, comprising only 0.6% of the total, which translates to roughly 12 customers, as illustrated in Figure 2.36. Upon closer examination, it was discovered that 90 customers do not belong to any of the recorded sources, suggesting the existence of additional channels for acquiring customers that are not captured in the current dataset. This insight underscores the importance of thorough data collection and analysis to ensure comprehensive understanding of customer acquisition channels. A new column *sourceOther* is created with 1 representing customers who do not belong to any of the 15 sources and 0 representing customers who belong to one of the 15 sources. *sourceOther* accounts for about 4.5% of the total.A graph of a number of numbers

Description automatically generated with medium confidence

**FIGURE 2.36** Customer distribution by sources

1. **PREDICTOR ANALYSIS AND RELEVANCY**

A correlation matrix as shown in Figure 3.1 was created to analyze the relationships between variables. Due to the large number of predictors, a simplified correlation matrix as shown in Figure 3.2 was generated by excluding predictors with low positive/negative correlations ( correlation value > .30).

A screen shot of a graph

Description automatically generated

**FIGURE 3.1** Correlation Matrix

**A graph with numbers and text

Description automatically generated with medium confidence**

**FIGURE 3.2** Simplified Correlation Matrix

A few expected correlations include Purchase and Spending, as well as firstUpdate and lastUpdate. Spending will only have a non-zero value if the customer is a purchaser. The linear relationship between firstUpdate and lastUpdate is due to the similar nature of the data they represent. Specifically, lastUpdate is always less than or equal to the first update, as illustrated in Figure 3.3.

A diagram of a number of dots

Description automatically generated

**FIGURE 3.3** Scatter plot of firstUpdate vs lastUpdate

Frequency shows a strong positive correlation with both outcome variables, Purchase and Spending, and a negative correlation with lastUpdate as shown in Figure 3.4. The positive correlation between frequency and purchase, is evident due to purchasers making at least one software purchase. However, it's noteworthy that non-purchasers also exhibit frequencies of purchase. This anomaly arises because the frequency attribute originates from the original data and is recorded within the past year by any of the firms in the consortium. Notably, purchasing amounts tend to increase with higher frequencies of purchases. The negative correlation between frequency and lastUpdate indicates that as the frequency of interactions with the system increases, the duration since the last interaction tends to decrease, suggesting more recent engagements with the platform or service.

A graph with numbers and text

Description automatically generated with medium confidence

**FIGURE 3.4** Distribution of frequency vs purchase, spending and lastUpdate.

Additionally, there is a strong correlation between sourceH and isResAddr. most of the customers acquired through source have provided their residential address as illustrated in Figure 3.5. A strong negative correlation is observed between sourceW and both firstUpdate and lastUpdate as seen in Figure 3.6. indicating that customers acquired through source have been added and updated recently.

A graph of a bar

Description automatically generated

**FIGURE 3.5** Distribution of source vs isResAddr.

A graph of data with numbers and a number of data

Description automatically generated with medium confidence

**FIGURE 3.6** Distribution of sourceW vs firstUpdate and lastUpdate.

The key predictors identified from the correlation matrices were frequency, lastUpdate, firstUpdate, isResAddr, source and sourceR. However, predictors such as sources and USCustomer may also be relevant.

1. **DIMENSION REDUCTION**

The variable "SequenceNo" was removed from the dataset as it lacks any meaningful information beyond providing uniqueness to the data. All other variables remain retained at this stage of the analysis. Given the high correlation between the "Spending" and "Purchase" variables, indicating redundancy, the "Spending" variable will be excluded from classification models, while the "Purchase" variable will be excluded from prediction models.

The number of observations (2000) in the dataset exceeds the number of features (26), hence the Variance Inflation Factor (VIF) is used to assess multicollinearity among variables. It was discovered that the VIF for lastUpdate, firstUpdate, and sourceW exceeds 4, indicating that these variables may offer redundant information, although not very serious, for predicting the outcome.

Although Principal Component Analysis (PCA) is useful for dimension reduction, it is not preferred in this case for several reasons. PCA works well when the number of observations (rows) is significantly larger than the number of variables (columns). With only 2000 records, using PCA might not be ideal, especially if there are many variables. PCA creates linear combinations of the original variables, making it challenging to interpret the resulting components. In a relatively small dataset such as this, interpretability is crucial to understand relationships between variables and their contributions to outcomes. Instead, techniques like correlation analysis or linear regression might be more appropriate. Additionally, PCA assumes that variables are linearly related and that there are no outliers. However, in this dataset, there were outliers in columns like spending and frequency. Therefore, a deliberate decision has been made not to use PCA for dimension reduction. Instead, methods like decision trees and stepwise regression will be employed during modeling to consider statistically significant attributes.

1. **DATA PARTITIONING**

In comparing the performance of different models, it's crucial to ensure that the chosen model can generalize well beyond the dataset it was trained on. To address this, the concept of data partitioning is employed, aiming to avoid overfitting. This involves dividing the data into subsets, each serving a specific purpose in the model evaluation process. Typically, two or three partitions are used: a training set, a validation set, and sometimes a holdout set.

In the current dataset, consisting of 2000 records, the data is randomly partitioned into training (800 records), validation (700 records), and holdout (500 records) sets. The training partition, being the largest, is used to build and train various models under consideration. These models are subsequently evaluated using the validation partition, which mimics the deployment scenario by applying the models to data they haven't encountered during training. When evaluating multiple models using the validation data and selecting the one that performs best, another aspect of the overfitting problem arises. Sometimes, certain aspects of the validation data coincidentally match the chosen model better than they do with other models. In simpler terms, when using the validation data to pick one model out of many, the chosen model may appear to perform better on the validation data than it would in reality. Applying the chosen model on the holdout dataset, which it hasn't encountered previously, will provide an unbiased estimate of its performance with new data.

When dealing with a small dataset, traditional data partitioning may not be advisable due to the risk of having too few records in each partition for effective model building and evaluation. Additionally, certain machine learning methods can be sensitive to slight variations in the training data, leading to different outcomes with different partitioning. In such cases, cross-validation, specifically k-fold cross-validation, becomes a valuable alternative. This procedure involves dividing the data into non-overlapping folds, often using k=5, where each fold represents 20% of the observations. The model is then fitted k times, using one fold as the validation set and the remaining k - 1 folds as the training set in each iteration. This ensures that each fold serves as the validation set at least once, generating predictions for every observation in the dataset. Combining the model's predictions from each validation set allows for an overall assessment of the model's performance.

For this project, data partitioning is chosen as the preferred approach to balance model building and evaluation with the available dataset of 2000 records. Given the moderate size of the dataset, the random partitioning into training, validation, and holdout sets ensures a reasonable representation of the data for effective model development and unbiased performance evaluation.

1. **DATA NORMALIZATION**

Normalization is a crucial step in data preprocessing, and its necessity depends on the characteristics of the variables. When variables share common units of measurement and their scales accurately reflect their importance, normalization may not be necessary. However, in cases where variables are measured in different units, making it challenging to compare their variability, normalization becomes essential. For instance, in the current dataset sample, binary variables like *USCustomer, isResAddr, webOrder, gender, sources,* and *purchase* coexist with variables like *lastUpdate* and *firstUpdate* measured in days, frequency represented as counts, and spending denoted in dollars. Here, the differing units make it unclear how to assess the significance of each variable's contribution, warranting normalization.

Moreover, variables like *firstUpdate* and *lastUpdate*, which represent days, could disproportionately influence the data due to their magnitudes, potentially skewing the mean. To mitigate this bias and ensure fair treatment of all variables during model training, validation, and testing, normalization techniques such as z-score scaling are applied. Z-score scaling helps to standardize the variables and bring them to a comparable scale, enabling more effective analysis and model performance evaluation across different types of variables.

Once the data is split into training, validation, and testing sets, the training dataset is normalized using z-scores. The mean and standard deviation of the training dataset are employed to normalize the validation and testing datasets. By applying the same normalization parameters to all datasets, consistency is ensured in the way variables are treated across the entire modeling process, from training to testing.

1. **CLASSIFICATION MODEL SELECTION**

Achieving the primary analytical goal of classifying customers as purchasers and non-purchasers, while also ranking them based on their probabilities of being purchasers, necessitates the application of classification algorithms. Among these, logistic regression and naive Bayes algorithms stand out as promising options due to their reliance on probabilities for classification, thus facilitating streamlined customer ranking. In contrast, algorithms such as classification trees or kNN prioritize classification based on majority votes and similarity metrics, respectively, making them less suitable for ranking objectives.

Initially, a **Logistic Regression Model** was trained on the training data. Employing stepwise regression, significant predictors were identified, encompassing channels of customer acquisition (*sourceA, sourceH, sourceU, sourceW, sourceR*), mode of order placement (*webOrder*), frequency of orders, and address type (*isResAddr*), all of which exert notable influence on the spending outcome variable. Subsequently, the model was refined using these predictors and evaluated on validation data. With a probability threshold set to 0.5, the model demonstrated an accuracy of 79.86% and a sensitivity of 75.07% in identifying purchasers.

Subsequently, a **Naïve Bayes Model** was trained on the same dataset using the significant predictors obtained from stepwise regression. Evaluation on the validation dataset revealed an accuracy of 73.86% and a sensitivity of 68%, inferior to the Logistic Regression Model's performance. Comparison of confusion matrices indicates Logistic Regression's superior accuracy and sensitivity in correctly classifying purchasers, as illustrated in Figure 7.1.

A comparison of a comparison of a model

Description automatically generated with medium confidence

**FIGURE 7.1** Confusion Matrices

Further analysis using cumulative gain and lift charts (Figures 7.2 and 7.3) underscores the Logistic Regression Model's efficacy in outperforming the base model and Naïve Bayes. Cumulative gains and lift charts are a graphical representation of the advantage of using a predictive model to choose which customers to select. The gains chart for Logistic regression shows that about 50% of the purchasers will be correctly identified from for 25% of the customers which is twice than that of Baseline which can only correctly identify 255 purchasers. The gain chart for naïve baye depicts that the model and the baseline performance is almost the same. The Naïve Bayes model's performance discrepancy can be attributed to its assumption of conditional independence among features, which, while simplifying the model, may not accurately reflect real-world scenarios. As a result of this assumption, the conditional probabilities in Naive Bayes models aren't guaranteed to sum up to 1, meaning they aren’t true probabilities and therefore cannot be used to rank customer correctly. Hence, this model is ruled out.

A graph of a graph of a number of cases

Description automatically generated with medium confidence

**FIGURE 7.2** Cumulative gain Chart Comparisons

A graph of a logistic regression

Description automatically generated with medium confidence

**FIGURE 7.3** Cumulative Lift Chart Comparisons

To refine purchaser identification, the Logistic Regression Model was optimized with an error cost ratio of 5:1, resulting in a markedly improved sensitivity of **97.99%,** with only 7 out of 349 purchasers misclassified. This enhanced model not only achieves robust classification but also enables customer ranking based on their probabilities of being purchasers, thereby enhancing decision-making processes. Given these considerations, **Logistic Regression emerges as the preferred choice for classification and ranking** tasks over Naïve Bayes.

1. **REGRESSION MODEL SELECTION**

After identifying potential purchasers using a classification model, the next step in the analysis is to predict the spending amount by these purchasers (not on non-purchasers). Two regression models, Multiple Linear Regression and Regression Trees are considered for this task. This report compares the performance and suitability of these models for predicting customer spending.

**Multiple Linear Regression** is chosen for its simplicity and ability to capture linear relationships between predictor variables and the outcome variable. After implementing stepwise regression on the training data, predictors such as *frequency, isResAddr, USCustomer, and lastUpdate* were found to be statistically significant. The model was fitted to the validation data, and the residuals were analyzed, and a histogram (in Figure 8.1) revealed a distribution closely resembling a normal distribution. The majority of records clustered around the mean, indicating a generally accurate prediction. However, a slight right skew was observed, suggesting a small number of predictions with significantly large discrepancies between the actual and predicted values. The correlation coefficient of 0.6906 indicates a strong positive relationship between the actual and predicted values, suggesting a good predictive capability of the model. However, the Mean Absolute Error (MAE) of 74.25 reveals that, on average, the model's predictions deviate from the observed spending by approximately $74.25, which is significantly lower than the deviation that would result from simply predicting the mean(93.90) for every instance.

A graph of a number of numbers

Description automatically generated

**FIGURE 8.1** Residual Distribution for Multiple Linear Regression

**Regression Trees** are chosen for their ability to capture complex interactions and non-linear relationships in the data. The model automatically detects important predictor variables and interactions, making it suitable for predicting customer spending behavior. After fitting the model to the validation set, a positive correlation between actual and predicted values was observed (correlation value of 0.6588). The MAE of 71.69 suggests that, on average, the model's predictions deviate from observed spending by approximately $71.69. Compared to predicting the mean spending value, the regression tree model offers significant improvement, indicating its effectiveness in predicting customer *spending.isResAddr, lastUpdate, and firstUpdate* are of significance in predicting the spending amount as shown in Figure 8.2.

A diagram of a company's flowchart

Description automatically generated

**FIGURE 8.2** Regression tree

Although the Regression Tree model has a slightly lower MAE, its inherent instability and complexity in interpreting relationships between predictors and the outcome variable favor the Multiple Linear Regression model. Despite its simplicity, the Multiple Linear Regression model performs reasonably well in predicting customer spending, making it the preferred choice among the two models.

1. **DATA SAMPLING**

When preparing data for modeling, a subset of records is often used due to limitations in computing power and algorithm capabilities. Accurate models can still be developed with fewer records, typically several thousand samples. However, in cases where the event being predicted is rare, like customer purchases in response to a mailing which is only 1065 customers out of 20,000 roughly accounting to 5.3%, randomly sampling records may result in too few instances of the event, leaving us with little information. Having plenty of data on non-purchasers but not enough on purchasers makes it hard to build a model that distinguishes between them. In such cases, it's important to give more weight to the rare class (purchasers) in our sampling so that we have a good mix of purchasers and non-purchasers in our sample. Hence, a sample of 2000 was extracted from the 20,000, with 1000 purchasers and 1000 non-purchasers.

Ensuring we have enough purchasers to train the model is essential, but it's also crucial to consider the costs of misclassification. When response rates are very low, it's more important to correctly identify purchasers than non-purchasers. Failing to catch purchasers can be costlier than spending extra time reviewing non-purchaser transactions.

The balanced dataset has a response rate of 50%, and we need to account for this discrepancy in response rates. To adjust for this, we must scale down the purchase rate predictions obtained from our model by multiplying each "case's probability of purchase" by the ratio of the observed response rate (5.3%) to the expected response rate (50%), which is approximately 0.106. This adjustment ensures that our predictions align with the actual response rates observed in the larger dataset, providing more accurate insights into customer behavior.

1. **PREDICTION**

The final phase of our analysis is to make predictions on the holdout dataset, which plays a pivotal role in assessing the generalization performance of our predictive models. To start, a classification tree model is utilized to classify customers in the holdout set based on their likelihood of making a purchase. The predicted probabilities of purchase are then computed and adjusted to account for the historical conversion rate. This step is crucial as it allows us to identify potential purchasers among the holdout dataset, enabling targeted marketing efforts towards individuals more likely to respond positively to our campaign.

The confusion matrix in Figure 10.1 reveals that our classification model achieved an overall accuracy of 73.2%. The sensitivity, or true positive rate, is high at 97.08%, indicating the model's effectiveness in identifying purchasers. However, the specificity, or true negative rate, is 51.15%, suggesting room for improvement in identifying non-purchasers. Despite this, the positive predictive value indicates that the model's predictions of purchases are generally accurate, which is crucial for directing marketing efforts towards potential purchasers.

A table with numbers and a number of objects

Description automatically generated with medium confidence

**FIGURE 10.1** Prediction Confusion Matrix

Next, a regression tree model is employed to predict the spending behavior of customers in the holdout set. By leveraging the features available in the dataset, such as customer demographics and past interactions, the model generates predictions on the expected spending amount for each customer. This information is vital for estimating the potential revenue generated from the campaign. The regression tree model applied to the holdout dataset yielded a correlation coefficient of 0.8129 when comparing predicted spending to actual spending, indicating a strong positive relationship between the predicted and observed values. The Mean Absolute Error (MAE) between the predicted and actual spending is 79.13. The average spending in the holdout dataset is $110.94, with spending ranging from $0 to $1500. Overall, the regression tree model demonstrates a relatively strong correlation with the actual spending, but there is still some error in the predictions.

The Cumulative Gains Chart in Figure 10.2 illustrates the performance of a predictive model. The blue line indicates that the model successfully prioritizes cases with higher expected spending. Specifically, when the top 100 cases are selected based on the model's predictions, the cumulative expected spending reaches approximately 20,000, and this figure nearly doubles to around 40,000 when the top 200 cases are targeted. The red dashed line represents a baseline of random targeting, which the model consistently outperforms, as evidenced by the gap between the lines. The chart shows diminishing returns after the top 200 cases, suggesting that the model is particularly effective at identifying the most lucrative cases early in the selection process. Overall, the model provides a clear advantage in selecting cases that are likely to yield higher spending compared to a random approach.

A graph with a line

Description automatically generated

**FIGURE 10.2** Cumulative gain chart based on prediction.

1. **CONCLUSION**

The firm's decision on customer selection methods depends on financial outcomes. If the firm opts to randomly select 180,000 customers from the pool, they anticipate approximately 9,540 responses, given a response rate of 5.3%. This expected response would generate a revenue of around $979,042.50, factoring in the average spending value of 102.62 derived from data. With mailing costs totaling $360,000 for the 180,000 customers at $2 per booklet, the estimated gross profit from this random selection approach stands at $619,042.50 after accounting for expenses. However, when using logistic regression to select the 180,000 customers from a larger pool of five million, the expected sales revenue is calculated using the average expected spending of from the model, resulting in a higher gross profit of $1,630,902. Based on this analysis, it is recommended that the client company use the logistic regression model to select customers for mailing and Regression Tree modelling to predict the sales and gross profit, to yield a significantly higher gross profit compared to random selection. The data-driven, model-based strategy will allow the firm to more effectively target customers who are likely to spend more, thereby optimizing their marketing expenditure and maximizing profitability.

1. **FUTURE WORK**

A few areas of research for the next quarter:

1. **Strategies for Worldwide Customer Base Expansion:**

Explore opportunities to diversify the customer base beyond the US market by devising targeted marketing strategies aimed at attracting customers from other countries. Advocate for the inclusion of firms with larger customer bases from diverse geographical locations within the consortium. This collaborative effort could facilitate North-Point's expansion into new international markets.

1. **Product Recommendations Enhancement:**

Enhance data collection efforts to gather more comprehensive information on customer purchasing patterns, including product preferences and frequency of purchases. Implement association rule mining techniques to analyze customer purchasing behavior and recommend complementary products, thereby enhancing the customer shopping experience and potentially increasing sales revenue.