CSDA 6010 Data Analytics Practicum

Project 1: North-Point Software Production Company

Sreeja Reddy Singidi

Webster University

Spring 2024

Instructor: Dr. Ali Ovlia

**TABLE OF CONTENTS**

[1. Introduction 5](#_Toc165756197)

[2. Project overview 5](#_Toc165756198)

[3. Business understanding 6](#_Toc165756199)

[4. Business goals 6](#_Toc165756200)

[5. Attributes definition 6](#_Toc165756201)

[6. Data understanding 8](#_Toc165756202)

[6.1 Structure 8](#_Toc165756203)

[6.2 Dimension 8](#_Toc165756204)

[6.3 Checking for missing values 8](#_Toc165756205)

[6.4 Checking for zero values 8](#_Toc165756206)

[6.5 Attribute analysis 8](#_Toc165756207)

[6.5.1 Distribution of outcome variables 9](#_Toc165756208)

[6.5.2 Distribution of categorical variables 10](#_Toc165756209)

[6.5.3 Distribution of numeric attributes 11](#_Toc165756210)

[6.5.4 Distribution of all sources when purchased 12](#_Toc165756211)

[6.5.5 Analyzing the variable frequency with outcome variables 12](#_Toc165756212)

[6.5.6 Average spending by gender 13](#_Toc165756213)

[6.5.7 Boxplots for numeric attribute 13](#_Toc165756214)

[7 Predictor analysis and relevancy 14](#_Toc165756215)

[8 Data transformations 15](#_Toc165756216)

[9 Dimension reduction 15](#_Toc165756217)

[10 Business considerations 15](#_Toc165756218)

[11 Data partitioning 16](#_Toc165756219)

[11.1 Why is data partitioning needed? 16](#_Toc165756220)

[11.2 Methods 16](#_Toc165756221)

[11.3 Importance of training, validation and testing 17](#_Toc165756222)

[12 Models selection: 17](#_Toc165756223)

[12.1 Goal 1- classification: 17](#_Toc165756224)

[12.1.1 Logistic regression 17](#_Toc165756225)

[12.1.2 Backward method 18](#_Toc165756226)

[12.1.3 Classification tree 19](#_Toc165756227)

[12.1.4 K-nearest neighbor 21](#_Toc165756228)

[Compare the classification models with accuracy and sensitivity 21](#_Toc165756229)

[12.2 Goal 2- regression 21](#_Toc165756230)

[12.2.1 Linear with stepwise regression 22](#_Toc165756231)

[12.2.2 Regression trees 22](#_Toc165756232)

[12.2.3 Comparing models for predicting spending: 23](#_Toc165756233)

[12.3 Models selected to predict on holdout data: 23](#_Toc165756234)

[13 Holdout data analysis and adjustments: 23](#_Toc165756235)

[13.1 Comparing logistic regression on validation and test data 23](#_Toc165756236)

[13.2 Comparing linear regression on validation and test data 23](#_Toc165756237)

[14 Profit analysis: 23](#_Toc165756238)

[14.1 Gross profit 24](#_Toc165756239)

[14.2 Predicted probability of purchase 24](#_Toc165756240)

[14.3 Adjusted probability of purchase 24](#_Toc165756241)

[14.4 Predicted spending value 24](#_Toc165756242)

[14.5 Expected spending 24](#_Toc165756243)

[14.6 Cumulative gain chart 25](#_Toc165756244)

[15 Conclusion 25](#_Toc165756245)

[16 Business recommendations 25](#_Toc165756246)

[Executive Summary: 26](#_Toc165756247)

# **LIST OF FIGURES**

Figure 1: Overview. 5

Figure 2: Distribution of Outcome Variable 9

Figure 3: Distribution of categorical variables. 10

Figure 4: Distribution of numeric attributes. 11

Figure 5: Distribution of numeric attributes. 12

Figure 6: Distribution of Purchase with Frequency 12

Figure 7: Distribution of Spending with Frequency 13

Figure 8: Average Spending by Gender 13

Figure 9: Boxplots for numeric attribute. 14

Figure 10: Scatterplot Matrix 14

Figure 11: Classification tree with all predictors. 20

Figure 12: Classification tree with important predictors. 20

Figure 13: Regression Tree 22

Figure 14: Cumulative gain chart. 25

# **LIST OF TABLES**

[Table 1: Attributes 7](#_Toc165756055)

[Table 2: New Column Names 15](#_Toc165756056)

[Table 3: Logistic Regression Models 19](#_Toc165756057)

[Table 4: Comparing models for predicting spending. 23](#_Toc165756058)

[Table 5: Comparing logistic regression on validation and test data. 23](#_Toc165756059)

[Table 6: Comparing linear regression with backward on validation and test data. 23](#_Toc165756060)

# INTRODUCTION

Initially, a software company, North-Point, sells games and educational software. This company has joined a consortium for customer name pooling. Predictive models will be developed in the project using historical data that focuses on customer response and spending patterns to improve the mailing list. The purpose is to improve the mailing plan for the next software release. In order to maximize the efficiency of models, a sample with an equal set of purchasers and non-purchasers will be used i.e. 1000 each- a total of 2000 observations in the dataset. Two models will be developed where one will categorize the customers into purchasers, and the other model will predict how much they have spent.

Outcome variables:

Purchase – It shows whether a customer purchased a product and responded to a test mailing.

Spending – This attribute comes into the picture if the customer has purchased something. If purchased, then how much they spent to buy that item is spending.

# PROJECT OVERVIEW

A diagram of a company

Description automatically generated

Figure 1: Overview.

# BUSINESS UNDERSTANDING

Companies in the consortium can trade lists of names to a pooled list of customers. Each company adds a list to the pool and can withdraw the same no. of names every quarter. Predictive modeling is done on the records (in the pool) to improve the selection of customer names. To enhance its mailing approach, North Point tested a list including 20,000 names out of a total of 5,000,000 names. 1,065 participants in this test purchased something, generating a 5.3% response rate.

They decided to use 1,000 purchasers and 1,000 non-purchasers to produce a balanced sample for analysis. This simplifies the dataset by producing a noticeable response rate of 50% or 0.5.

By multiplying everyone’s expected probability of purchasing by 0.106, the adjustment brings the dataset's purchase rate back to its initial 5.3%. By doing this, it is ensured that the models are optimized according to the real response rate that was noted during the first mailing test.

# BUSINESS GOALS

The objective is to maximize possible gross profit while minimizing mailing expenses. If the company chooses at random from the pool of 180,000 customers, it wants to calculate the gross profit it could expect. Each list booklet costs about $2 to mail, which covers the cost of printing, postage, and other mailing expenses.

The final goal is to pick customers who tend to purchase and spend more to boost the response rate and income from mailing campaigns.

# ATTRIBUTES DEFINITION

1. sequence\_number: It is the unique identification of each record that is used to refer to records. This attribute might not have any direct impact on prediction.
2. US: If a customer is residing in the US the record displays 1 otherwise 0. When developing a model, it will be useful for predictions based on location.
3. "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": A customer has been drawn from these sources. This is a binary attribute where 1 indicates a customer name is taken from that particular channel. It will have an impact on customer behavior while making purchases.
4. Freq: It is a numeric attribute that shows no. of transactions in the last year. Spending can be predicted and shows how active a customer is in purchasing a product.
5. last\_update\_days\_ago: Numeric attribute showing how many no. of days ago the last update was made.
6. X1st\_update\_days\_ago: Numeric attribute showing how many no. of days ago the first update was made.
7. Web.order: Binary attribute that distinguishes customers based on how they make a purchase. If a customer purchased through the web, then 1 otherwise 0.
8. Gender.male: Gender attribute- if male 1 otherwise 0.
9. Address\_is\_res: Describe whether the address is the same as the residential address. 1 for residential address and 0 for non-residential address.
10. Purchase: This is a target variable for classification, which shows whether a customer made a purchase and responded to test mailings.
11. Spending: This is a target variable for regression that helps maximize gross profit by identifying high-spending customers and offers insights into their purchasing habits.

|  |  |  |  |
| --- | --- | --- | --- |
| Column No. | Variable Name | Description | Type |
| 1 | sequence number | A unique identification number is assigned to each record | Numerical |
| 2 | US | Indicates whether the customer resides in the US (1 for yes, 0 for no) | Binary |
| 3-17 | source\_\* | Binary attributes indicating the customer's source channel which has 15 channels | Binary |
| 18 | Freq | Number of transactions made by the customer in the last year. | Numerical |
| 19 | last\_update\_days\_ago | Number of days since the last update was made. | Numerical |
| 20 | X1st\_update\_days\_ago | Number of days since the first update was made. | Numerical |
| 21 | Web.order | Indicates whether the customer made a purchase through the web (1 for yes, 0 for no) | Binary |
| 22 | Gender.male | Gender of the customer (1 for male, 0 for female) | Binary |
| 23 | Address\_is\_res | Address is same as residential or not (1 for yes, 0 for no) | Binary |
| 24 | Purchase | The customer made purchase in test mailing. | Binary |
| 25 | Spending | Amount in dollars customer spent in test mailing. | Numerical |

Table 1: Attributes

# DATA UNDERSTANDING

## STRUCTURE

Deciding whether attributes in the North-Point project are binary, numerical, or categorical is made easier by understanding the structure.

## DIMENSION

Understanding the dataset's dimensions—that is, number of rows and columns—is crucial in evaluating the project's scope.

## CHECKING FOR MISSING VALUES

Model predictions may be skewed or incorrect if there are missing values. For, suppose important information is missing then algorithms might not function properly. Complete data provides better modeling in predicting spending and purchaser classification. Checking for missing values can be performed by using is.na () function. This data set does not contain any missing values.

## CHECKING FOR ZERO VALUES

For checking 0 values, consider numeric data. This dataset's numerical columns are spending, frequency, last update, and first update. Freq has 398 "0" values, meaning the client did not make any purchases in the previous year. The dataset indicates that there are 1000 non-purchasers. Still, the 0 values are only 999 since one record (sequence no. 711) in the dataset indicates that a customer has not made any purchases and has only paid $1.

## ATTRIBUTE ANALYSIS

Analyzing attributes is a crucial step to understanding the features of each variable in the dataset. This involves determining the data types, attribute distribution, identifying outliers, and correlation between the variables.

* + 1. **DISTRIBUTION OF OUTCOME VARIABLES**

Purchase: The outcome variable purchase is divided into 2 categories- purchasers and non-purchasers equally as seen below.

Spending: The graph is skewed right where the maximum amount a customer spends is $1500. Most of the people spent below 500 to purchase something.

|  |  |
| --- | --- |
| A green and blue rectangular boxes  Description automatically generated | A purple and white graph  Description automatically generated |

Figure 2: Distribution of Outcome Variable

* + 1. **DISTRIBUTION OF CATEGORICAL VARIABLES**

We can conclude from these distributions that most of the customers are from the US who tend to buy through the web. Gender does not show much of a difference, but male customers are slightly high compared to other genders.

A screenshot of a graph

Description automatically generated

Figure 3: Distribution of categorical variables.

### DISTRIBUTION OF NUMERIC ATTRIBUTES

Frequency and spending are right skewed as seen. Customers who made 0,1 transactions are high, so it tops the chart. In the same manner, customers are more likely to spend up to 500 and the maximum amount spent is 1500. Most of the people whose first and last update is around 2500-3000 days ago.

A group of yellow bars

Description automatically generated

Figure 4: Distribution of numeric attributes.

### DISTRIBUTION OF ALL SOURCES WHEN PURCHASED

The Figure 4, shows the count of each channel or source when purchasing something.

We can see that ‘source a’ has more customers who have purchased and replied to mailing.

A graph of colored bars

Description automatically generated with medium confidence

Figure 5: Distribution of numeric attributes.

### ANALYZING THE VARIABLE FREQUENCY WITH OUTCOME VARIABLES

Purchase:

A graph of a graph showing different colored bars

Description automatically generated with medium confidence

Figure 6: Distribution of Purchase with Frequency

**Spending:**

A graph of a chart

Description automatically generated with medium confidence

Figure 7: Distribution of Spending with Frequency

Plotted the bar graphs as seen from Figure 5 and 6, to show the relationship between purchase and frequency. From these plots, we could analyze that customers who order once make more purchases.

The scatter plot shows the relationship between the attributes Freq and Spending and it is evident that if the no. of transactions is less the spending is also low. And if the transactions are more customers spend more.

### AVERAGE SPENDING BY GENDER

By taking the average of spending, other genders have spent more compared to males.

A green and blue squares

Description automatically generated

Figure 8: Average Spending by Gender

### BOXPLOTS FOR NUMERIC ATTRIBUTE

For all numeric attributes boxplot can show the outliers. But, in this case, the outliers do not come into consideration because high spending can show customers who purchased with high amounts rather than errors.

A diagram of a box plot

Description automatically generated

Figure 9: Boxplots for numeric attribute.

* + 1. SCATTERPLOT MATRIX TO KNOW THE RELATION BETWEEN EACH NUMERIC ATTRIBUTE

This entire graph shows the correlation between all the numeric attributes with their distributions using pairs.panels() function. The last updated and the first updated show a high correlation of 0.81. and the least relation between frequency and spending is noticed.

A screenshot of a graph

Description automatically generated

Figure 10: Scatterplot Matrix

# **PREDICTOR ANALYSIS AND RELEVANCY**

**Sequence numbers** serve just like an index of records which does not give any information that can be used in the models further. This can be considered as an **irrelevant attribute** and removed before building the models for predictions which can actually simplify the dataset. For knowing the important predictors, logistic regression with stepwise can be used which will be performed in the further steps.

There are **90 rows** where all the **source attributes** are marked **0**. Take all those sources with 0 values into **a different set** to analyze. When analyzing this data, records with sources having all 0 values have purchases and spending. It is assumed that these customers are either directly connected with the company or are part of its own database that understands the behavior of customers who interact with the company directly without being taken from any other channel.

# **DATA TRANSFORMATIONS**

The dataset is easier to read when column names are clear and easy to read. The readers can easily understand each attribute's significance without referring to the document. It is simpler to understand the meaning of each attribute if its name is given properly. In this dataset, a few column names can be changed.

|  |  |
| --- | --- |
| Existing column names | New column names |
| Freq | Frequency |
| last\_update\_days\_ago | Days\_since\_Last\_Update |
| X1st\_update\_days\_ago’ | Days\_since\_First\_Update |
| Web.order’ | online\_order |
| gender\_male | gender\_male |

Table 2: New Column Names

# **DIMENSION REDUCTION**

The goal of dimension reduction is to keep the important information in a dataset while reducing the number of variables. This can be advantageous for making the model simpler, lowering the chance of overfitting, and increasing computational effectiveness.

**Correlation analysis:**

This can be performed on numeric data by using summary statistics to identify the highly correlated variables and remove them. The last update and first update as seen in the matrix are near 1 which is highly correlated and overlaps the information, so we can remove anyone but after logistic regression stepwise the comparison and reduction if needed can be done.

Mostly the predictors that do not contribute much to the model can be known after building it. So, we can remove the predictors then and evaluate the performances for both models from which we can compare the performances.

After building the logistic regression model with stepwise to select important predictors, will come back to this step and make the necessary changes if it is required to be removed.

# **BUSINESS CONSIDERATIONS**

To improve its marketing activities, the company uses tools from financial analysis and predictive modeling. Predictive models are created by analyzing previous customer responses and spending data to enhance the process of choosing names for mailing campaigns. Using these models, the business can improve response rates, more precisely target its customer base, and maximize its marketing budget.

To guarantee precise modeling and analysis, North-Point generates a balanced sample dataset with an equal proportion of buyers and non-buyers. When dealing with uneven response rates, in particular, this balanced sample makes the dataset easier to understand and allows for more precise forecasts.

To guarantee that prediction models accurately represent real-world circumstances, modifications are implemented to the dataset to preserve the initial response rate recorded during test mailings.

The current budget is adequate to meet the project's needs, eliminating the necessity for additional resources currently.

* *Note - Reduction using models:*

Dimension reduction is considered by important attributes in the dataset. When building the model, logistic regression can be used for this purpose. If the model is performing better with the important predictors – only those can be taken into consideration and suppose there is no improvement, then can consider all the predictors. In that case, dimension reduction will not take place. Not every dataset needs to perform dimension reduction. Based on the insights, whatever is needed can be done.

# **DATA PARTITIONING**

## **WHY IS DATA PARTITIONING NEEDED?**

Data partitioning is important because it evaluates model performance on data that is not used in the training set. This guarantees that the model applies effectively to the new set of data. Model performance can be improved without overfitting. Overfitting is avoided by dividing the data or splitting it. Data partitioning is an essential step to ensure that models are reliable and able to produce correct predictions.

## **METHODS**

Holdout Method: This is splitting the data into 2 or 3 parts training and testing or training, validation, and testing. This works by building the model on training, if validation is considered then predict the model on validation and evaluate different model performances which is called as fine-tuning, and then test on completely new data with the best model from validation.

Cross-Validation: The dataset is divided into folds. The model will be trained on selected folds like k-1 folds and tested on other folds.

To achieve effective model performance and interpretability in the classification process, logistic regression with backward selection will be used to check which model is performing better.

As a part of the project requirement data partitioning is done this way because it serves the purpose:

Training with 800 records

Validation with 700 records

Testing with 500 records

## **IMPORTANCE OF TRAINING, VALIDATION AND TESTING**

Its main purpose is before the final assessment on the test data, it optimizes the model on validation data. The model is guaranteed to be trained on a training set, validated to adjust parameters and prevent overfitting (validation set), and then select the best model from validation, then predict the selected-on test data.

After partitioning the data, models should be selected for classifying a customer into purchaser or non-purchaser & to predict the spending when a purchase is made. The best model should be selected by comparing the performance of different models based on evaluation metrics. Training the model, evaluating, comparing, and selecting the best will be seen in further steps.

# **MODELS SELECTION:**

## **GOAL 1- CLASSIFICATION:**

The main goal as discussed above is to first classify a customer into a purchaser or non-purchaser. This enables companies to target those who are most likely to react favorably to campaigns, promotions, and mailing. This can be achieved by performing logistic regression and with the backward method as per the requirements. Although other possible models can also be built for just comparing purposes.

As we are aware the company is involved in mailing campaigns, classification models pick customers who respond to the campaign, which helps with mailing list efficiency and raises response rates while lowering mailing costs. This step is essential for keeping current customers as well as for attracting new ones. Spending can be removed in this classification.

**To classify the customer into a purchaser or non-purchaser these models can be used:**

### **LOGISTIC REGRESSION**

This glm() model is well suited for binary classification problems that give probabilities of a customer belonging to a particular class.

Initially consider all the predictors and build the model on a train set, then make predictions on the validation data. It gives predicted probabilities of purchase for each observation in the validation dataset. The probabilities are continuous values between 0 and 1. But to classify a customer it should be binary like 0- for non-purchasers and 1 for purchasers. So, to convert these probabilities into binary values threshold is used. The threshold is 0.5. If the probabilities are greater than the threshold- classified as 1(purchaser) less than the threshold- classified as 0(non-purchaser). Binary labels should be factored, and levels should be set to “1” and “0”. By doing so the prediction will be appropriate.

### **BACKWARD METHOD**

To improve the model performance other methods like selecting important predictors can be taken into consideration and checking if there is any improvement. StepAIC will be available in the MASS package.

This method gives the most relevant and significant variables for predicting whether a customer is a purchaser or non-purchaser. By identifying and eliminating strongly correlated variables, stepwise techniques can lessen multicollinearity problems that could compromise the model's stability and interpretability.

Mostly backward stepwise is preferable, so for this model backward is used.

After performing stepwise on the initial model the predictors, it is considering important are*:*

*source\_a, source\_e, source\_h , source\_r , source\_s, source\_t, source\_u, source\_p, source\_x, source\_w, Frequency, Days\_since\_last\_update, online\_order, Address\_is\_res*

In the earlier steps, dimension reduction shows a high correlation between the last update and the first update where it is mentioned that anyone can be removed. From the important predictors, we can see that the first update is removed as it is considered as irrelevant. Model evaluation should be for every model to conclude which model is performing well and justify it.

For this model and the model with all predictors, there is not much difference but a slight decrease in sensitivity and accuracy in model 1 is noticed.

**Comparing Logistic Regression models based on accuracy and sensitivity***:*

After performing the model with stepwise and all predictors, can check by removing insignificant variables in each step. By doing so if there is any improvement in the model performance on the validation data we can select that model to predict on holdout data.

Model 2: source\_s amd source\_x are insignificant so these variables are improved in this model.

Model 3: source\_t is insignificant so removing that in this model.

Model 4: source\_e and last\_updaate\_days\_ago are less significant marked as ‘.’, removing them.

Model 5: source\_p is insignificant so removing that in this model.

Model 6: source\_r has just one star so trying by removing it and checking the performance.

|  |  |  |
| --- | --- | --- |
| Logistic Regression Model | Accuracy | Sensitivity |
| Initial Model | 79.57% | 77.59% |
| Model 1(After stepwise) | 79.29% | 76.44% |
| Model 2 (Significant variables set 1) | 79.29% | 76.15% |
| Model 3 (Significant variables set 2) | 78.86% | 74.71% |
| Model 4 (Significant variables set 3) | 79.71% | 75% |
| Model 5 (Significant variables set 4) | 79.57% | 74.71% |
| Model 6 (Significant variables set 5) | 78.57% | 72.99% |

Table 3: Logistic Regression Models

Model 4 seems to have the highest accuracy compared to all other models, but the sensitivity is just 75%. Our target is to pick the best customers for the firm, which means classifying as a purchaser is important so the sensitivity shows how likely a customer can be classified as a purchaser correctly. By considering that fact we must go with more sensitivity so an initial model with all predictors can be considered from all these models.

### **CLASSIFICATION TREE**

The classification tree is easy to understand and gives a simple way to demonstrate decision rules. Each node represents a decision. When it comes to predicting the target variable, it chooses the important predictors by itself. By this, we can easily identify the important attributes that impact the classification. The leaf nodes will have 0 or 1 showing non-purchasers or purchasers, predictor and value of the split will be shown right below the node. The count of the classes will be shown inside the node.

The training dataset builds the decision tree based on the partition. Another set of 700 records which is not used in training is used to check the performance- notices for overfitting where they do not perform well on new data.

To compare the tree - important predictors are also taken into consideration to check the performance. The process is the same as the above. Here when considering the important predictors model tends to perform well compared to another tree model.

**Tree with all predictors:**

A diagram of a computer

Description automatically generated

Figure 11: Classification tree with all predictors.

**Tree using important predictors:**

A diagram of a computer network

Description automatically generated

Figure 12: Classification tree with important predictors.

### **K-NEAREST NEIGHBOR**

Problems involving binary and multi-class classification can be handled by k-NN. It is easy to apply k-NN to the North Point dataset, where we classify customers into purchasers and non-purchasers. Based on the majority class of the k nearest neighbors to a particular data point, KNN generates predictions.

**KNN with all predictors:**

The model is trained. Next predictions are made using the model on the validation dataset. Following the prediction- accuracy and sensitivity are considered for the model performance. Validation data reveals a sensitivity of 76.15.

**KNN with important predictors from the backward method:**

A set of predictors is taken from the backward method performed in the logistic regression to compare the performances. Here, the goal is to determine which model performs better- only significant predictors in a model or by using all the predictors.

**COMPARE THE CLASSIFICATION MODELS WITH ACCURACY AND SENSITIVITY**

|  |  |  |
| --- | --- | --- |
| Models | Accuracy | Sensitivity |
| Logistic Regression model with all predictors | 79.57% | 77.59% |
| Logistic Regression model with backward | 79.29% | 76.44% |
| Classification Tree | 76.86% | 67.53% |
| Classification Tree (with important predictors) | 77.5% | 75.86% |
| KNN (k=1) | 80.71% | 76.15% |
| KNN (with important predictors) | 78.86% | 77.30% |

Table 3: Classification model selection

The classification tree has the least sensitivity so that model can be eliminated, considering Logistic regression with all predictors to perform on hold-out data as the sensitivity is more for that model and meets the project requirements.

**Logistic regression with all predictors can be considered to perform predictions on the final set which is the test dataset.**

## **GOAL 2- REGRESSION**

The goal is to project the amount of money that customers will likely spend on their purchases. With the aid of this projection, North Point will be able to precisely project income and customize marketing tactics to optimize profitability***.***

**To predict spending value for purchasers these models can be used*:***

### **LINEAR WITH STEPWISE REGRESSION**

Linear regression is used to predict numeric outcomes. This data predicts the spending should contain only purchasers which means Purchasers = 1 should be considered. Later, the purchase should be removed while building the model.

Based on the most influential predictors, they can make marketing strategies. This is a fundamental model for predicting spending. A useful way to verify the validity and reliability of a model is to compare its performance with sophisticated models like regression trees.

The first step is to build a multiple linear regression model including all predictors. Build a model on the train data and evaluate using the validation set. Now, perform stepAIC with backward selection to get the important predictors. Once we have the important predictors, repeat the same steps and compare the results. This model will be using mean absolute error for evaluation. The least mae shows better performance in the model.

Comparing both the linear regression models, MAE was low for the backward method (109).

So, let's dive into exploring regression trees for comparison.

### **REGRESSION TREES**

The regression tree helps in predicting spending values for each purchaser. We must first identify predictors and later split the data. The model is built on training data. Tree splits based on the highest standard deviation reduction (SDR).

The tree automatically takes the important predictors into consideration so build the tree on the train and evaluate the mae score on validation data.

A diagram of a number

Description automatically generated

Figure 13: Regression Tree

### **COMPARING MODELS FOR PREDICTING SPENDING:**

|  |  |
| --- | --- |
| Models | Evaluation (MAE) |
| Multiple linear regression with all predictors | 111.902 |
| Multiple linear regression after stepwise | 109.8421 |
| Regression tree | 100.581 |

Table 4: Comparing models for predicting spending.

**Linear regression after stepwise can be considered as it deals with important predictors only that have low MAE which helps to deal with model complexity, also it is a part of the specifications mentioned.**

## **MODELS SELECTED TO PREDICT ON HOLDOUT DATA:**

* Classification - Logistic regression with all predictors
* Regression - Linear regression using backward.

# **HOLDOUT DATA ANALYSIS AND ADJUSTMENTS:**

It's crucial to examine the original holdout data and make the required modifications for more precise predictions after developing the selected models and assessing their effectiveness. By doing this step, we can be sure that the models are realistic and optimized for real-world situations.

After predicting the selected models on test data these are the results:

Evaluation of test data for both selected models shows improved performance.

## **COMPARING LOGISTIC REGRESSION ON VALIDATION AND TEST DATA**

|  |  |  |
| --- | --- | --- |
|  | ACCURACY | SENSITIVITY |
| VALIDATION DATA | 79.57% | 77.59% |
| TEST DATA | 82.6% | 79% |

Table 5: Comparing logistic regression on validation and test data.

## **COMPARING LINEAR REGRESSION ON VALIDATION AND TEST DATA**

|  |  |
| --- | --- |
|  | MAE (Mean Absolute Error) |
| VALIDATION DATA | 109 |
| TEST DATA | 113 |

Table 6: Comparing linear regression with backward on validation and test data.

# **PROFIT ANALYSIS:**

North Point's business goals include maximizing gross profit, reducing mailing costs, and successfully targeting customers to increase response rates and revenue from mailing campaigns can be done by adding the following columns:

## **GROSS PROFIT**

North Point can more accurately estimate and plan its financial resources by estimating the possible gross profit. It assists in overall financial planning for marketing efforts and will know where it requires improvement. The overall goal is to maximize the profits.

Gross Profit can be calculated by considering:

* Total no. of customer in mailing pool = 180,000
* Response Rate = 0.053 (5.3%)
* Average of spending when purchase is 1 = 205
* Each mailing cost = $2

By multiplying the response rate (0.053) by the average spending of the customers who made a purchase and deducting the mailing cost per customer ($2). This gives an approximate idea of how much money North-Point Software Production Company might make from the campaign. Estimated gross profit is $1598075.

## **PREDICTED PROBABILITY OF PURCHASE**

Based on the predictive logistic regression model on test data, this column assists North Point in determining the probability that each customer will make a purchase. It enables the business to provide preference to clients who are more likely to react favorably to marketing initiatives, which improves resource allocation and boosts conversion rates.

## **ADJUSTED PROBABILITY OF PURCHASE**

The Adjusted Probability of Purchase column adjusts for purchaser oversampling, guaranteeing that the predictive models correctly represent the actual response rate seen in the first mailing test. North Point can produce more accurate forecasts and adjust marketing initiatives by aligning the dataset with the real purchase distribution. This column can be added by multiplying the predicted probability of purchase by the original purchase rate which is 0.1065.

## **PREDICTED SPENDING VALUE**

North Point may more precisely predict revenue and customize marketing tactics to increase profitability by projecting the amount of money that customers are likely to spend on purchases. Personalized offers and targeted marketing efforts boost client engagement and loyalty by considering the spending patterns of various customer categories.

## **EXPECTED SPENDING**

The expected spending column offers a more accurate estimate of customer’s expenses. With this modification, oversampling is taken into consideration and accurate projections are used to provide revenue forecasts. This allows North Point to make well-informed judgments.

## **CUMULATIVE GAIN CHART**

The cumulative gain chart provides information on the efficiency of targeting methods by visualizing cumulative predicted spending as a function of records targeted. It assists North Point in determining high-value customer segments, assessing the return on investment of various marketing strategies, and optimizing campaign effectiveness to increase revenue production. The chart shows selecting around 400 customers from the test records will maximize the spending which increases the profit.

A graph and a graph

Description automatically generated

Figure 14: Cumulative gain chart.

# **CONCLUSION**

North-Point successfully implemented predictive modeling to optimize its mailing campaigns, resulting in improved response rates and increased profitability. The adjusted dataset ensured realistic modeling, while financial analysis tools aided in profit estimation and resource allocation. Overall, North-Point's strategic approach led to enhanced marketing effectiveness and maximized gross profit, validating the project's objectives and methodologies.

# **BUSINESS RECOMMENDATIONS**

North Point must constantly assess how well its marketing initiatives and prediction models are working and modify its approach as necessary. North Point can preserve a competitive edge and be successful in its marketing efforts over the long run by remaining adaptable and sensitive to shifts in customer behavior and market dynamics. Using the cumulative chart, North Point should constantly analyze the success of its targeting methods to pinpoint high value.

# EXECUTIVE SUMMARY:

**NAME:** Sreeja Reddy Singidi

**DATE:** 04-24-2024

**OPPORTUNITIES:**

The project provided valuable insights into customer behavior and spending patterns, enabling North-Point to better understand its audience and tailor its offerings to meet their needs. This deeper understanding facilitated more personalized marketing strategies, driving increased customer satisfaction and loyalty.

**SOLUTIONS:**

The company tried to improve its mailing strategy for games and educational software by utilizing financial research and predictive modeling. Additionally, throughout the project, significant attention was paid to business factors like optimizing gross profit, reducing mailing costs, and successfully targeting customers. More informed decisions and resource allocation were made possible by the inclusion of modified probabilities, spending forecasts, and predicted purchase probabilities.