STA 6543: Predictive Modeing

Project: Fundraising Write-up

Name:

Angel Manuel Perez

1. Business Objectives and Goals:

Business Objective: Increase the effectiveness of a mailing campaign by targeting individuals likely to donate to a cause.

Goals:

1. Identify individuals who are more likely to donate.
2. Optimize the allocation of resources for the mailing campaign.
3. Maximize the return on investment (ROI) for the campaign.

2. Data Sources and Data Used:

Data Sources: The dataset used for analysis was obtained from [source]. It contains information about individuals, including demographic data and past donation history.

Reasoning for Weighted Sampling: Weighted sampling was used to produce a training set with equal numbers of donors and non-donors to address the issue of class imbalance. This ensures that the model is trained on a balanced dataset, which can lead to better performance and generalization. Simple random sampling from the original dataset could have resulted in a training set with unequal class distribution, which may bias the model towards the majority class.

3. Type of Analysis Performed:

Analysis Performed: Two machine learning models, Random Forest Classifier and Gradient Boosting Classifier, were trained on the dataset to predict the likelihood of donation based on individual characteristics.

Methodology:

- Data Preprocessing: Encoding categorical variables, splitting the dataset into training and validation sets.

- Model Training: Random Forest Classifier and Gradient Boosting Classifier were trained on the training set.
- Model Evaluation: Performance metrics such as accuracy, precision, recall, F1-score, ROC AUC were calculated on the validation set.
- Visualization: ROC curves and Precision-Recall curves were plotted to assess model performance.

4. Exclusions:

No exclusions were made in this analysis.

5. Variable Transformations:

Categorical variables were encoded using one-hot encoding to convert them into numerical format for model training. No other variable transformations were performed.

6. Business Inputs:

The primary business inputs for this analysis include:

- Human resources for data collection and analysis.
- Financial resources for model development and implementation.
- Information about past donation history and individual characteristics.

7. Methodology Used, Background, Benefits:

The methodology used in this analysis involves machine learning techniques to predict donor behavior. Random Forest Classifier and Gradient Boosting Classifier were chosen for their ability to handle complex relationships in the data and provide accurate predictions. The benefits of this approach include improved targeting of potential donors, leading to higher campaign effectiveness and ROI.

8. Model Performance and Validation Results:

Random Forest Classifier:

- Accuracy: 0.567
- Precision: 0.573
- Recall: 0.554
- F1-score: 0.564
- ROC AUC: 0.567

Gradient Boosting Classifier:

- Accuracy: 0.553

- Precision: 0.558
- Recall: 0.554
- F1-score: 0.556
- ROC AUC: 0.553

Both models achieved similar performance metrics, with Random Forest Classifier slightly outperforming Gradient Boosting Classifier in terms of accuracy and ROC AUC.

9. Cut-Off Analysis:

No specific cut-off analysis was performed in this analysis.

10. Recommendations:

Determining the percentage of data to allocate for a mailing campaign involves balancing several factors to achieve the most effective outcome. Here are some recommendations to consider when deciding on the percentage of data for the campaign:

1. **Response Rate Analysis**: Analyze past mailing campaign data to understand the response rate. Determine the proportion of recipients who responded positively to the campaign in the past. This can serve as a baseline for estimating the expected response rate for the current campaign.

2. **Cost-Benefit Analysis**: Evaluate the cost associated with the mailing campaign, including production and postage costs, against the potential benefits such as donations or customer acquisitions. Calculate the expected return on investment (ROI) based on historical data and campaign goals.

3. **Segmentation Strategy**: Segment the donor database based on factors such as past donation history, demographics, or behavioral attributes. Allocate a higher percentage of data to segments that have shown higher response rates in the past or are more likely to generate significant returns.

4. **Testing and Optimization**: Allocate a portion of the data (e.g., 10-20%) for testing different campaign strategies, messaging, or creative elements. Use A/B testing or multivariate testing to identify the most effective approach before rolling out the campaign to the entire dataset.

5. **Risk Mitigation**: Consider allocating a conservative percentage of data initially to minimize the risk associated with the campaign. Gradually increase the percentage based on the performance of early mailings and response rates.

6. **Resource Constraints**: Take into account any limitations or constraints such as budget, manpower, or production capacity. Ensure that the allocated percentage of data is manageable within the available resources.

Based on these considerations, I would recommend initially allocating around 10-20% of the dataset for the mailing campaign. This allows for testing different strategies, minimizing risk, and optimizing the campaign before scaling up to a larger percentage of data. As the campaign progresses and performance metrics are monitored, adjustments can be made to the allocation percentage to maximize effectiveness and ROI. Based on the analysis results, it is recommended to use the Random Forest Classifier model for the mailing campaign, as it demonstrated slightly better performance compared to the Gradient Boosting Classifier. Additionally, it is recommended to allocate resources to target individuals who are predicted to have a higher likelihood of donation, as identified by the model.

11. Pseudo Codes for Implementation:

```r
# R code for implementing the Random Forest Classifier model

# Load required libraries
library(randomForest)

# Load and preprocess the data
data <- read.csv("data.csv")
# Perform data preprocessing steps (e.g., encoding categorical variables)

# Split the data into training and validation sets
set.seed(12345)
train_index <- sample(1:nrow(data), 0.8*nrow(data))
train_data <- data[train_index, ]
val_data <- data[-train_index, ]

# Train the Random Forest Classifier model
model <- randomForest(target_variable ~ ., data = train_data, ntree = 100,
mtry = 3)

# Make predictions on the validation set
predictions <- predict(model, newdata = val_data)

# Evaluate model performance
# Calculate accuracy, precision, recall, F1-score, ROC AUC, etc.

# Plot ROC curve and Precision-Recall curve
# Plot ROC curve
plot(roc_curve_data)

# Plot Precision-Recall curve
plot(precision_recall_curve_data)
```

This pseudo code outlines the steps for implementing the Random Forest Classifier model in R, including data preprocessing, model training, prediction, evaluation, and visualization of results. Similar steps can be followed for the Gradient Boosting Classifier model.