*FINAL PROJECT REPORT*

*A data analysis project to provide business understanding to a Non-Profit Organization using advanced modeling techniques.*

|  |  |  |
| --- | --- | --- |
| S/N | Names | Student ID |
| 1 | Chibuzor Chukwu | 11693062 |

**Tables of Content**

|  |  |
| --- | --- |
| Items | Page number |
| Introduction | 3 |
| Data Understanding/EDA | 5 |
| Data Wrangling/Preparation | 10 |
| Data Modeling | 10 |
| Model Evaluation | 15 |
| Model Deployment | 16 |
| Conclusion | 17 |
| References | 18 |

**Introduction**

A Non-profit organization (NPO) is a legal entity organized and operated for a collective, public, or social benefit, as opposed to an entity that operates a business aiming to generate profit for its owners (NPO Wikipedia 2024). According to National Center for Charitable Statistics (NCCS), there are over 1.5 million nonprofit organizations registered in the United States. According to National Council of Nonprofits, every person in the United States benefits from the work of nonprofits in one way or another whether they realize it or not.

The goals and activities of non-profit organizations are mostly dependent on donations. Efforts for direct marketing are essential for attracting contributors and collecting donations. However, if not effectively targeted, these efforts may result in high expenses and poor results. The cost-effectiveness of marketing activities may be significantly increased by identifying donor behaviors and estimating future donors.

This project aims to maximize the profit/revenue generated by our client’s non-profit organization. According to our client’s records they had an estimated overall response rate of 10% from their donor’s contribution in their most recent marketing campaign. Our analytics team will be building classification models that will capture future donors. We will also develop predictive model to predict expected gift amounts from our likely donors.

**Business Understanding**

NPO are not motivated by generation profit, however they must bring in enough revenue to pursue their social goals. We will be utilizing modeling techniques on the data set provided by an NPO to better understand more about the characteristics and behaviors of their donors to maximize their profits and increase revenue.

The dataset includes socioeconomic, demographic, and donation-related data such as neighborhood indicators, homeownership, income, gender, and area. These details offer us a more complex picture of the backgrounds of possible contributors. Furthermore, factors that affect donor behavior, such number of donations, value of gifts, and promotions received, provide insight into patterns of donating and involvement. Two response variables are included, which are crucial: "DAMT" for estimating donation amounts and "DONR" for donor categorization.

**Analytics Questions**

1. How do numerous socioeconomic factors or geographic locations affect the likelihood of donations?
2. Is predictive modeling suitable to predict future contribution amounts and classifications based on historical data?
3. In what ways might the results of the predictive models be applied to improve the performance of direct marketing campaigns?
4. What strategies might be developed using the findings from the model to better determine potential contributors and allocate resources?

**Data Understanding/EDA**

Data Description:

Our dataset which in turn consists of the past donor activities has a total of around 6002 rows and

|  |  |  |
| --- | --- | --- |
| Column | Description | Type |
| ID | Unique identifier for each donor. | Alphanumeric |
| Region | Geographic region of the donor. | Categorical |
| Ownd | Homeownership status. | Binary |
| Kids | Number of children. | Numeric |
| Inc | Individual income. | Numeric |
| Sex | Gender of the donor. | Categorical |
| Wlth | Wealth rating. | Numeric |
| Hv | Home value. | Numeric |
| Incmed | Median income in the donor's area. | Numeric |
| Incavg | Average income in the donor's area. | Numeric |
| Low | Indicator of low-income status. | Binary |
| Npro | Non-profit engagement level. | Numeric |
| Gifdol | Total dollar amount of gifts donated. | Numeric |
| Gifl | Largest single gift amount. | Numeric |
| Gifr | Most recent gift amounts. | Numeric |
| Mdon | Median donation amount. | Numeric |
| Lag | Time since last gift. | Numeric |
| Gifa | Average gift amount. | Numeric |
| Donr | Donor status. | Binary |
| Damt | Amount donated in the last campaign. | Numeric |

**Data Cleaning and Missing Values**

However, there are no missing values in this dataset, making it a complete collection of data. Every one of the 6002 rows in our dataset has data for each of the mentioned columns. This suggests that correction and handling of missing values are not required.   
We may use the whole dataset and have confidence that incomplete or incorrect information won't affect our research if there are no missing values in the dataset. We may confidently continue our study since every observation gives us a full picture of the variables we are examining.A screenshot of a computer

Description automatically generated

**Descriptive Statistics:**

In our thorough examination of the dataset, we explored a wide range of variables that provided useful insights regarding donor behavior and characteristics. The dataset consisting of 6002 observations, had no missing values, which ensured a strong basis for our analysis. Now, we will dive deep into the descriptive nature of our important variables.

**Target Variables:**

* Donor Status (DONR): 50.12% of donors were identified as '0' (non-donors) and 49.88% as '1' (donors), making 'donr' the target variable for our classification model with an equally divided distribution.
* Amount Donated in the Last Campaign (DAMT):  The mean donation for 'damt,' the target variable in our regression model, was $7.21, with a standard deviation of $7.36. A distribution that was close to normal but had a peak below a typical normal distribution was suggested by the nearly-zero skewness (-0.11) and negative kurtosis (-1.83).

These figures offer an advanced view on our dataset, highlighting the variety and complexity of the variables affecting donor behavior. This thorough investigation provides a strong basis for our further modeling efforts.

**Input Variables:**

* Income (INC): With 7 different income levels, level 4 was the most common income level among participants, representing 45.88% of the dataset. Five was the second most common income level, representing 14.80% of the data. The individuals who contributed appear to have an important level of economic variety based on this distribution.
* Homeownership (OWND): 88.45% of donors were homeowners (shown by 1), whereas 11.55% did not own a home (shown by 0). That is a huge percentage. The bulk of the participants may have secure financial situations, as seen by the high ratio of homeowners.
* Region: Five distinct geographical areas received donations from the people. With 34.71% of the dataset, the most common area was 'ter2', followed by 'ter1' with 20.14%. This implies a donor base that is geographically varied.
* Gender (SEX): 39.22% of donors were female (marked by 0) and 60.78% of donors were male (indicated by 1). This dataset revealed an average gender imbalance.
* Wealth Rating (WLTH): Out of 10 levels, level 8 (38.55%) had the highest frequency of the wealth rating variable, followed by level 9 (27.09%). This suggests that those who give have the tendency to rate their wealth higher.
* Average Gift Amount (GIFA): With a standard deviation of $6.53 and a mean of $11.68, this variable exhibited a broad range of donations amounts, suggesting donor goodwill movements. A distribution with a peak and a right skewness (1.74) and kurtosis (5.91) are suggested.

A screenshot of a computer

Description automatically generated

* Total Dollar Amount of Gifts (GIFDOL): The mean gift amount was $115.80, but there was an important amount of variability and a concentration of lesser values, as indicated by the high standard deviation of $86.54 and an extreme skewness of 6.09.
* Largest Single Gift Amount (GIFL): This variable displayed an almost right-skewed distribution (skewness of 7.18), indicating that big donations were less common but varied greatly in value. It also exhibited a mean of $22.98 and a high standard deviation of $29.40.
* Most Recent Gift Amount (GIFR), Home Value (hv), Income Average (incavg), Income Median (incmed): The skewness of these factors varied in different ranges and degrees, which is indicative of the members' recent participation and changing financial situation.
* Number of Children (KIDS), Time Since Last Gift (LAG), Low-Income Indicator (LOW), Median Donation Amount (MDON), Non-Profit Engagement Level (NPRO): These factors provided further information on the demographics and patterns of involvement of the donors.

**Data Wrangling/ Preparation**

Data wrangling (data cleaning) refers to a variety of processes designed to transform raw data into more readily used formats [Business Insights Blog, 2021]. We used the DMDB node under the explore tab to decide if it needed cleaning and there were no missing values found. The dataset has 20 columns representing different variables and 6002 rows representing the records for each observation. We made use of SAS Enterprise Miner for all our models both classification and prediction. We also used data partitioning and transforming when building some of our models.

**Data Modeling**

**Classification models**: We made use of 3 classification models and each model made use of the DONR variable as a target variable.

**Model 1 KNN**

KNN Model: For our KNN model, the data set was divided into 2 parts training. We made use of the data partition node in SAS Enterprise Miner to split our dataset into training and validation sets using the 70-30 rule. We also used the transform variable node to balance the weights in our variables (some variables have a wider range than others, and this creates imbalance). The MBR (Memory-Based reasoning) node was used for our KNN models and evaluated the values of k =1 to k =12. The setting in the train menu for each model was set to k + 1 for the number of neighbors.

A screenshot of a computer

Description automatically generated

The Model comparison node was used to select the best model. MBR 12 has a misclassification rate of 0.16537 which gives 16.5% and made it the best model. We will be using this model for our evaluation.

**Model 2 Naïve Bayes model**

Naïve Bayes model was used as another one of our classification models. The data partition node was used to split our dataset using the 70-30 rule. The DONR variable was made our target variable, we used the HP BN classifier connected to the partition node to execute the naïve bayes model.

**A screenshot of a computer

Description automatically generated**A screenshot of a computer

Description automatically generated

After evaluating our model, we got a misclassification of 0.10 as seen in our results. We also got an ASE of 0.08.

**Model 3 Logistic Regression model**

LR model was conducted for classification purpose and the data was imported to our diagram using import node. We Identified the relevant target and predictors variables for our logistic regression model. We also divided the data set into training and validation sets using data partition node whereas the data set allocation is applied to 70% for training and 30% for validation. Regression node has been implemented by activating the below equations: two-factor interaction (Activation of two factor interaction might enhance the model capability to encounter complex relationship and increase accurate predictability), and polynomial terms (Activation polynomial terms allow the model to encounter no-linear relationship between predictors and target variables) and setting the regression type to logistic and model selection to stepwise.

The Classification chart below shows how our model correctly classified if an individual was donor or non-donor.

**A screenshot of a computer

Description automatically generated**A screenshot of a computer screen

Description automatically generated

Our MISC rate has a rate of 12%

**Predictive Models:** We made use of 3 prediction models using the DAMT variable as the target variable.

**Model 1 KNN**

KNN Model: Using the earlier setting where we evaluated the values of K =1 to K = 12, we make use of KNN model as a predictive model. In this model of KNN we used DAMT as our target variable instead of DONR. We also normalized the predictors as we did earlier by setting range standardization on all intervals input except DAMT because it is the target variable.

A screenshot of a computer

Description automatically generated

The Model comparison node was used to select the best model. MBR 11 had an ASE of 0.035175 which made it the best model. We will be using this model for our evaluation.

**Model 2 Random Forest**

The HPForest node under the HPDM tab was used on our dataset. The variables INC, KIDS, REGION, WITH and OWND were the top 5 from our variable importance table which signifies how important they are when predicting the donation amount (DAMT). We got an Average Squared Error of 20.23.

A screenshot of a computer

Description automatically generated

**Model 3 Gradient Boosting**

We made use of the gradient boosting node under the model tab. The gradient boosting predictive model helps us identify the most important predictors when predicting the donation amount (DAMT). KIDS, INC, REGION, OWND, and WITH were the top 5 variables with the most importance. We can also see that the gradient boosting model shares similar results with our random forest model. We got and Average Squared Error of 21.69.

A screenshot of a computer

Description automatically generated

**Model Evaluation**

Focusing on the business objective for the non-profit organization of our clients. We are to develop a classification model using the dataset from their most recent campaign that will effectively show all future donors. This will help us to increase the donor recipients who will receive flyers and reduce non donor recipients from receiving flyers.

We are also expected to build a predictive model to predict the expected gift amounts from our future donors. Our clients have an average donation of $14.5 with an expected response rate of 10% and each mailing costs $2 to produce. Further calculations show that our clients are running at a loss of $0.55 for each mailing.

We will be using different model assessment criteria to evaluate our validation data results and determine the best model before deploying it.

|  |  |  |  |
| --- | --- | --- | --- |
| Classification Model | Misclassification rate  (MISC) | Sensitivity | Average Squared Error |
| KNN | 16.5% | 91% | 0.126 |
| Logistic Regression | 12% | 92% | 0.09 |
| Naïve Bayes | 10% | 94% | 0.08 |

*Table 1 above shows the comparison between our classification models based on MISC, Sensitivity, Specificity and Average Squared Error.*

Sensitivity = (TP/TP+FN)

KNN = (822/822+77) = 0.91

Logistic regression = (825/825+74) = 0.92

Naïve Bayes = (849/849+50) = 0.94

|  |  |
| --- | --- |
| Predictive Model | Average Squared Error |
| KNN | 0.035 |
| Random Forests | 20.23 |
| Gradient Boosting | 21.69 |

*Table 2 abive shows the comparison between our predictive models based on Average Squared Error*

**Model Deployment**

All models were deployed using SAS Enterprise Miner. We will be determining the profit with our classification model results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FN | TN | FP | TP | Misclassification Rate |
| KNN | 77 | 682 | 221 | 822 | 16.5% |
| Naïve Bayes | 50 | 779 | 124 | 849 | 10% |
| Log Regression | 74 | 806 | 97 | 825 | 12% |

Profits: (TP\* Mail donation) – (TP+FP) \* Mailing cost

KNN = {(822\*14.50) – (822+221) \*2} = $9,833

Naïve Bayes = {(849\*14.50) – (849+124) \*2} = $10,364.5

Logistic regression = {(825\*14.50) – (825+97) \*2} = $10,118.50

The Naïve Bayes model will produce the best profits with an amount of $10,364.5. This means for us to maximize profit for our clients we will be using the Naïve bayes model. We were also able to provide a score data for our best classification and prediction model using the score node and the nonprofit score dataset.

**Business recommendations for client**

Rather than mass marketing approach, our models show the necessary individual wo are more likely to be donors. We were able to measure the level of our classification models using sensitivity, Naïve Bayes model has the best sensitivity which means it will be the best approach to capture people who are most likely to be donors. Our predictor models also show variables that need to be considered when trying to increase donation amount. Kids, household income, regions, wealth ratings and household owners need to be considered when trying to increase our donation amount.

**Conclusion**

The aim of project was to improve the cost effectiveness of a nonprofit organization mail outreach campaign. Key findings from our report suggest that Naïve Bayes model proved to be the best choice for resource allocation and project profitability. In addition, the Naïve Bayes model was the model profitable model with $10,364.5. Our KNN model was also used for both classification an prediction and it came out as the predictive model due to having the least ASE (Average Squared Error). We had minor limitations which included complexity of using SAS, Future research and enhancement areas include exploring more analytical tools like Python and R, new variables, external datasets, and more machine learning techniques to improve mode accuracy.

References

1. Nonprofit organization (2024) Wikipedia. Available at: https://en.wikipedia.org/wiki/Nonprofit\_organization (Accessed: 22 February 2024).
2. Nonprofit impact in communities (no date) National Council of Nonprofits. Available at: https://www.councilofnonprofits.org/about-americas-nonprofits/nonprofit-impact-communities (Accessed: 22 February 2024).
3. Data wrangling: What it is & why it’s important (2021) Business Insights Blog. Available at: https://online.hbs.edu/blog/post/data-wrangling (Accessed: 21 February 2024).
4. Choueiry, G. (2024). Why and When to Include Interactions in a Regression Model. QUANTIFYING HEALTH. <https://quantifyinghealth.com/why-and-when-to-include-interactions-in-a-regression-model/>
5. Frost, J. (n.d.). Mean Squared Error (MSE). Statistics By Jim. <https://statisticsbyjim.com/regression/mean-squared-error-mse/>
6. Tabrez, S. T. (2023, July 13). MSE as an evaluation metric for Regression Models. Tutorials Points.<https://www.tutorialspoint.com/mse-as-an-evaluation-metric-for-regressionmodels>
7. SAS (2023, January 30). Analyze with a Logistic Regression Model. SAS Help Center. <https://documentation.sas.com/doc/en/emgsj/15.3/p0tad07m88xmotn1c78zunl7sm5i.htm>
8. Z. (2022, March 25). *Misclassification Rate in Machine Learning: Definition & Example*. Statology.<https://www.statology.org/misclassificationrate/#:~:text=The%20value%20for%20misclassification%20rate%20can%20range%20from,to%20predict%20the%20outcomes%20of%20the%20response%20variable>.