University of North Texas

ADTA 5230.401 Data Analytics II

Final Project Report

Group 3:

Hemalatha Kamesh

Sonali Sabnam

Mandaichelle Casby

July 25, 2024, 11:59pm

**Executive Summary:**

**Background:**

Group 3 has been hired by a non-profit organization seeking to enhance the cost-effectiveness of their direct marketing campaigns targeted at donors. The organization's historical data indicates an average response rate of 10% to their mailings, with an average donation of $14.50 per responder. A cost of $2.00 per recipient, which includes a gift of personalized address labels. The current approach yields a loss of -$0.55 per mailing, demanding optimized recipient selection through predictive modeling.

**Business Requirement:** To develop a classification model based on the most recent campaign data to effectively capture likely donors and predict their expected gift amounts.

**Framework Followed:** CRoss Industry Standard Process for Data Mining (Crisp-DM)

Group 3 followed the Crisp-DM Framework to develop predictive models for improving the cost-effectiveness of direct marketing campaigns. The CRISP-DM framework provides a structured yet flexible approach to achieve business goals with data based iterative refinement, and clear communication of results and recommendations to the client.

A diagram of a process

Description automatically generated

**Figure:** CRISP-DM framework

Group 3 identified the business problem, performed EDA, built, trained and tested classification and prediction models to identify the donors, donation amount and underlying trends leading to donations.

**Team Contributions:**

* **Business/ Data Understanding and EDA:** Hemalatha Kamesh
* **Modeling and Evaluation:** Sonali Sabnam
* **Deployment and Conclusion:** Mandaichelle Casby

**Conclusion:** In conclusion, Group 3 utilized the Crisp-DM framework to answer the business questions and provide the organization with helpful insights to minimize the campaign cost and increase the efficiency per mailing with a predicted profit of **$10.7804.**

**Business Understanding / Analytics Question(s):**

**Business Understanding:**

We are Group 3, and we were approached by a nonprofit organization to develop predictive models to improve the cost-effectiveness of their direct marketing campaigns to donors at a lower cost. The response rate currently is 10% with each contributor roughly donating about $14.50 on average. With the current technique and response rate the organization is at a net loss of $0.55 with each time they send out the mail. Therefore, the organization has reached out to us so we can develop a classification model using data from the most recent campaign that can effectively capture likely donors and their expected donation amount.

**Analytics Question(s):**

**Question 1:** How can we identify who is most likely to reply to the mailing campaign and donate?

**Question 2:** What is the expected donation amount from the predicted donors?

**EDA & Data Understanding / Data Preparation:**

**EDA:**

The Exploratory Data Analysis (EDA) is a foundational step in the CRISP-DM framework that guided the entire data mining process by enabling Group 3 to gain a clear understanding of the dataset and identify key patterns and relationships that will drive future modeling and analysis. The EDA enabled Group 3 to formulate the analytics questions and select relevant variables based on data-driven insights, while identifying data quality issues such as outliers, to ensure the data was reliable and suitable for following modeling stages (**Source:** Han & Kamber, 2006).

**Data Understanding:**

A screenshot of a graph

Description automatically generated

The above correlation matrix shows that when there is a higher income level, it is strongly associated with higher home values. We can tell from this graph that households with more children are less likely to be donors and that people who rent homes are less likely to donate than homeowners. Also, past donation amounts correlate very highly with future donation amounts and are great predictors as a donor who has donated a lot in the past is more likely to donate a similar amount in the future.

A graph of wealth rating and home value

Description automatically generated

The above scatterplot shows that when home values increase there is also a higher wealth rating for both non-donor and donor groups. Also, the donors appear to have a higher average home value compared to non-donors at similar wealth ratings which also indicates that there is a positive correlation between home value and the likelihood of being a donor.

A graph with red and blue dots

Description automatically generated

In the scatterplot above, we can see that when wealth ratings increase the donation amounts also have an increase when we're looking at donors which suggests a positive correlation between wealth rating and donation amount. Additionally, non-donors show a zero-donation amount rate among all wealth ratings, which makes sense because non-donors do not donate. This trend suggests that wealthier individuals tend to donate more.

The histogram below depicts the frequency in which an amount is donated. Because half of the people in this survey did not donate there are many donations of with 0 value. But if we only consider those who did donate, we can see most people donated 14 dollars and donations range from 8-27 dollars. Most people who did donate donated about the same amount (with a slight right skew).

A graph of distribution of donation amount

Description automatically generated

The below histogram shows that the income category 4 average donation is slightly below 10 but taking into consideration about half of the participates are in this category, it is likely that this might influenced the data and showing the highest average donation comes from category 4.

A graph of a number of people

Description automatically generated with medium confidence

In the graph below it shows the average amount of money donated by each region. Ter2 has the highest donation amount with Ter1 coming in second.

A graph of a number of different colored bars

Description automatically generated

In the graph below it shows the homeowners donate about dollar 14 as a median and some of the donations even reach up to dollar 27. However, non-homeowners have a much narrower range of donations with most of them being around 10 − 15 and very few are much more than that.

A diagram of a donation amount

Description automatically generated

A graph of a graph of wealth rating

Description automatically generated with medium confidence

In the above boxplot we can see that increase in wealth ratings also there is also a higher maximum gift amount as we see more and larger outliers. The median, however, seems to remain constant among all wealth ratings, but wealthier individuals have the potential to donate larger gifts even though all the medians are approximately the same.

**Data Preparation:**

The data set provided by the client consisted of 20 columns with no missing values and without any duplicate rows and columns. Group 3 removed the ID column because it was not necessary for the analysis and added a new column called 'income\_metric' by averaging the 'incmed' and 'incavg' columns. Then Group 3 converted the categorical 'region' column into binary values and did not remove any outliers from the dataset because it has been oversampled to have equal proportions of donors and non-donors, so any outlier removal is unnecessary.

**Modeling:**

Modeling is the fourth step in the CRISP-DM (Cross Industry Standard Process for Data Mining) framework. Based on the problem statement we identified the Dependent variables “donr” and “damt” from the given data.

“**donr**”: As we must capture the likely donors, we select the variable “donr” from the given dataset to be our dependent variable for Classification Models.

“**damt**”: As we must predict the expected gift amount from donors we select “damt” to be the dependent variable from the given dataset for the Prediction models.

We explored various models to identify and selected the best suited models for our problem statement. The details are described below.

**Problem Statement:** To capture likely donors and predict the expected gift amount from donors

* **Classification**: To classify the record as “donor” or “non-donor”

**Dependent Variable (Y):** “donr”

**Independent Variables (X):** "region", "ownd", "kids", "inc", "sex", "wlth", "hv", "low", "npro", "gifdol", "gifl", "gifr", "mdon", "lag", "gifa",”income\_metric”

**Columns not considered as Independent Variables from the dataset:** ’id’, 'incavg', ‘incmed’, 'donr', 'damt'

**Models Considered for Classification:**

KNN, Neural network, Logistic Regression, Random Forest, Support Vector Machine

**Models Used for Classification:**

Logistic Regression, Random Forest, Support Vector Machine

* **Prediction:** To predict the amount of donation

**Dependent Variable (Y):** “damt”

**Independent Variable (X):** "region", "ownd", "kids", "inc", "sex", "wlth", "hv", "low", "npro", "gifdol", "gifl", "gifr", "mdon", "lag", "gifa",”income\_metric”

**Columns not considered as Independent Variables from the dataset:** ’id’, 'incavg', ‘incmed’, 'donr', 'damt'

**Models Considered for Prediction:** KNN, Neural network, Linear Regression, Random Forest Regressor, Gradient Boosting Regressor

**Models Used for Prediction:** Linear Regression, Random Forest Regressor, Gradient Boosting Regressor

**Reason for not using KNN:** As we are dealing with oversampled data, there is a fair chance of overfitting

**Reason for not using Neural Network:** It is a strong modelling algorithm, but it requires a large dataset for high accuracy.

**Modelling Process followed:** As we have two different datasets, we have used the below approach for both Classification and Regression.

* + **Training:** Use the entire training dataset.
  + **Preprocess test data:** Apply the same preprocessing steps as the training data using the fitted preprocessor.
  + **Use test data for making predictions:** Predict the target variable using the trained model on the test data.

**Evaluation the Modelling Results:**

**Classification:** Though all the three models perform fair, we can see that the Random Forest algorithm has the same accuracy as Support Vector Machine (highest accuracy) but the lowest misclassification rate. This makes it the best suited model.

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Misclassification Rate** |
| Logistic Regression | 0.84 | 0.15653622 |
| Random Forest | 0.89 | 0.105745212 |
| Support Vector Machine | 0.89 | 0.111573689 |

**Table:** Results of the Classification Models

The below screenshots are the code outputs for Logistic Regression, Random Forest and Support Vector Machine algorithms.

**A screenshot of a graph

Description automatically generated** **A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Figure:** Confusion Matrices for Classification Algorithms

**Regression:** The results below are recorded from the Regression models. We can see that Gradient Boosting Regressor has the lowest Mean Square Error (MSE) and highest R-Squared values which makes it the best model among the three models. The result screenshot is shown below.

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R^2** |
| Linear Regression | 1.70207119 | 0.555396394 |
| Random Forest Regressor | 1.525500501 | 0.601519004 |
| Gradient Boosting Regressor | 1.295824701 | 0.661513374 |

**Table:** Results of the Regression Models

A close-up of a number

Description automatically generated

**Figure:** Results for Regression Models

**Result:** Out of the 2007 records used for prediction, our best model (Random Forest) has captured 405 people as likely donors, and the donation amount goes from $ 10.01 to $19.33.

**Note:**

* We have used the “nonprofit.xlsx” for training and testing. “nonprofit\_score.xlsx” was used for prediction.
* The code file includes the implementation and results for KNN and Neural Network too. As we are not considering these models for our evaluation, the output file does not contain the predictions for KNN and Neural Network.
* As part of the project requirements, we have saved the classification and prediction output to a csv file and will be submitted. This output file contains the results columns as Pred\_Donr\_log\_reg, Pred\_Donr\_rf, Pred\_Donr\_svm, Pred\_Damt\_log\_reg, Pred\_Damt\_rf, Pred\_Damt\_gbr.
* As the Random Forest model has accuracy of 89% (very high), we are considering the output of Random Forest Classifier model to be the input to the prediction models.
* For the result we are considering Random Forest model for classification and Gradient Boosting Regressor for regression as described above.

**Deployment:**

**Findings and Relevant Business Insights:**

The following insights highlight the relationships between various factors like income, homeownership, home values, and donation behaviors which can be utilized to enhance the client’s targeted fundraising strategies and donor engagement efforts:

* We used the output of best classification and regression models to determine the factors impacting whether donations will be made by an individual and the predicted amount of the donations.
* Key Insights include:
  + Individuals with no children tend to donate more amount.
  + People who have donated a high amount in the past continue to be likely donors.
  + People having a home value greater than 100,000 dollars tend to be potential donors.
  + Homeowners donate more frequent and higher donation amounts compared to non-homeowners.

The below pairwise scatter plot shows the supporting evidence for the above-mentioned points.

**A group of graphs with different colored dots

Description automatically generated with medium confidence**

**Figure:** Pairwise scatter plot for identifying relationship between the key variables.

**Actionable Business Recommendations:**

Group 3 provides the client the following Actionable Business Recommendations, based on the insights gained from the developed models:

1. **Targeted Donor Acquisition:**

* A targeted approach where the predicted donors are be reached by the marketing campaign increases the chances of donation while optimizing resource allocation

1. **Personalized Donation Solicitation:**

* Based on the predicted donation amount, the organization can customize donation requests and plan to attain higher contribution from the likely high donating individuals. This can boost the donation revenue significantly.

1. **Campaign Strategy Optimization:**

* It is encouraged to modify the organization’s strategies to point to the identified donor segments such as high-income individuals, homeowners, individuals who do not have children. This can boost donor engagement and drive better outcome by generating more donations.

1. **Implementation of the model on the score data for managerial and non­technical audience:**

* The classification and prediction models can be easily integrated for new donor profiles. The output of the model evaluations is stored in csv files which are easily understandable by any managerial and non­technical audience.
* The graphs make it very easy to understand the donor trends and donation amount and identify the key segments to focus on for devising the donation campaign strategy

**Expected Profit (or Loss):**

Average predicted donation amount (Gradient Boosting Regressor): $14.36

Accuracy of classification (Random Forest Classifier): 0.89

Cost per mailing: $2

**Profit = (14.36\*0.89) – 2 = 10.7804**

With this calculation, we can expect a profit of **$10.7804** per mailing.

**Conclusion & Discussion:**

**Summary of highlights and findings:**

* We identified the business problem, performed exploratory data analysis and used it for modeling. We evaluated the models to identify the best model.
* We used the output of the best classification model (Random Forest Classifier) and prediction model (Gradient Boosting Regressor) to identify the potential donors and predict donation amount.
* We analyzed the output data to determine the potential segments and trends.
* We classified **405** individuals as donors and the average predicted donation amount is **$14.36.**
* The average profit per mailing is **$10.7804.**

**Limitations:**

* **Limited size of training data:** A bigger training dataset can help the model to learn all the variations efficiently.
* **Change in donor’s behavior:** Economic and social factors impacting the donor’s behavior.

**Reference List:**

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.