Dr. Quintar ADTA 5230 Fall 2023

Troy Krupinski-

Prashant Thapaliya

Gayathri Devi Bobbarapalli

Raghavireddy Kondam-11661742

Jashwanth Kalyan.

Predictive Modeling for Nonprofit Donations

# Appendix

1. **Introduction**
2. **Exploratory Data Analysis**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**
7. **Conclusion**

# 1.Introduction

The nonprofit organization is facing a big challenge in making their direct marketing to past donors more cost-effective. Recent data shows that about 10% of people respond to their campaigns, and those who do donate around $14.50 on average. However, each campaign costs $2.00 for making and sending, which leads to a loss of -$0.55 for each mailing. To tackle this issue, the organization wants to get better at predicting which people are likely to donate. They plan to use data from their latest campaign to build a smart model. The goal is to send the campaign to more people who are likely to donate and fewer to those who probably won't, aiming to improve the overall results.

The data set provided an opportunity to explore patterns within a nonprofit organization's donor database to predict future donations. The primary business question addressed was whether it's possible to accurately predict whether an individual would donate (DONR variable), and the potential donation amount (DAMT variable) based on demographic, financial, and donation history as well as predicting the best model for the project.

# 2.Exploratory Data Analysis | EDA

The first analysis included figuring out how the dataset was organized, finding any missing values, and evaluating the distribution of the data. The dataset includes historical donation statistics, income indicators, and a variety of demographic characteristics. Several columns had missing values, which were handled by imputing the mean of the numerical features. Histograms and count plots were two important visualizations that provided information about the distribution of categories and numerical variables. The data set provided an opportunity to explore patterns within a nonprofit organization's donor database to predict future donations. The primary business question addressed was whether it's possible to accurately predict whether an individual would donate (DONR variable), and the potential donation amount (DAMT variable) based on demographic, financial, and donation history.

**2.1The EDA analysis with Business questions**

1. How does the percentage of low income affect the likelihood of making a large donation?

A graph of a distribution of low income

Description automatically generated

The graph shows the distribution of donations by percentage of low income. The graph is right-skewed, which means that the majority of donations are relatively small, but there are a few exceptionally large donations.

Donations are more likely to be made by people who are lower in income than those who are higher in income.

1. Does the avg home value determine the amount of donations?

A graph of distribution of home value

Description automatically generated

The donation distribution depending on average home value is displayed in the graph. The graph indicates that most donations come from homeowners whose average property worth is between $200,000 and $300,000. Additionally, a sizable portion of donations come from individuals whose homes are worth between $300,000 and $400,000. People who own properties outside of or above this range, however, donate much less frequently.

1. What is the average donation amount for each income category?

A graph of a distribution of household income

Description automatically generated

The average donation amount varies by income category. Households in the $2,000-$2,500 income category make the largest average donations. Households in the $500-$1,000 income categories make the smallest average donations.

1. What are the important insights from Count plot of region?

A graph of different colored bars

Description automatically generated

The chart shows the number of donations by region. The regions with the highest number of donations are ter2, ter1and5 with over 1250 donations e. Ter3and ter4 have the fewest donations, with less than 1000 donations each.

1. Do you think families with more children are more likely or less likely to donate to charity? Why?

A graph of a number of children

Description automatically generated

The graph shows that the majority of families have 0 or 2 children, followed by families with 3 children. The number of families with 1 or more than 4 children decreased significantly. Families with 2 children may be more likely to have a sense of civic responsibility and feel a need to give back to their community, which could lead to higher donation amounts overall. Families with more than 2 children may have less disposable income, which could lead to lower donation amounts.

1. What are general Insights of the donation amount distribution?

A graph of distribution of donation

Description automatically generated

The graph shows the distribution of donation amounts. The donation amounts range from $0 to $3000, with the most common donation amount being between $10 and $20.

There is a long tail of larger donations, with a few donations over $1000.

1. What is the distribution of average family income for the donations?

A graph of a distribution of family income

Description automatically generated

The graph shows the distribution of donation amounts based on the average home value. The graph shows that the majority of donations are made by people who have homes with an average value of $30,000-$70,000.

However, the number of donations made by people with homes above or below this range decreases significantly.

1. What are the insights for the Average gift amount?

A graph of a distribution of gifts

Description automatically generated

The graph shows the distribution of average gift amount by donation count. The most common average gift amount is around $10, but it can be as low as $0 or as high as $400.

The graph is skewed to the right, meaning that there are more donations with a lower average gift amount than donations with a higher average gift amount. This is because there are many small donations, but there are also a few exceptionally large donations.

**2.2 The EDA plots**

A diagram of distribution of donor

Description automatically generatedA graph of distribution of homeowner

Description automatically generatedA graph of distribution of male and female

Description automatically generated

This graph is left skew and tells there are enormous number of females.

This graph shoes that there are more number of homeowners

This graph showes there are equal number of donators and non donators

A graph of a distribution of gifts

Description automatically generatedA graph of a gift

Description automatically generatedA graph of a distribution of lifetime promotion

Description automatically generated

This graph is right skew the count is decreasing as the lifetime promotions are increasing.

This graph is right skew and showes that lifetime gift amount are more in between 0- 250$

This graph is right skew and showes that large amount of gifts are betwe en 0- 100$

A graph of a distribution of family income

Description automatically generatedA graph of distribution of the last donation

Description automatically generatedA graph of a distribution of gifts

Description automatically generated

This graph shows a right skew the count is less for larger gift amount.

This graph shows that the past dontation are high from in15-25 months

The median family income is a right skew has more count in between 0-100.

.

A graph with a line going up

Description automatically generatedA graph of a distribution of gifts

Description automatically generated

.

.

with the count of more than 2000 wealth rating 8 are more. Has a left skew

There is more 5-to-10-month difference first and second gift.

# 3.Data Preparation

In order to prepare the data, it was necessary to remove unnecessary columns, deal with missing values (which were none), and use one-hot encoding to encode categorical variables. For the purposes of classification (DONR) and regression (DAMT), the dataset was divided into predictors (ID, DONR, and DAMT excluded) and target variables. Numerical features underwent standard scaling to get the dataset ready for modeling.

**Dropping Unnecessary Columns:**

The rationale behind excluding columns such as 'ID,' 'donr,' and 'damt' is to concentrate on pertinent features for the machine learning models. 'ID' typically serves as an identifier without contributing to the model's predictive capabilities. Meanwhile, 'donr' and 'damt,' being target variables for classification and regression, are omitted from the feature set to prevent information leakage.

**Handling Missing Values in Numeric Columns:**

To maintain dataset integrity, missing values in numeric columns are often filled through imputation. In this case, the mean imputation strategy is employed, replacing missing values with the mean of their respective columns. This method is prevalent when assuming that missing values occur randomly, and imputing the mean preserves the overall distribution.

**Separating Target Variables:**

The standard practice in supervised machine learning involves segregating target variables ('donr' and 'damt') from the feature set. This step enables the models to learn patterns from the features and make predictions regarding the target variables.

**Identifying Categorical Columns for One-Hot Encoding:**

Given that machine learning models typically require numerical input, categorical variables must be converted into a numerical format. One-hot encoding, a technique where categorical variables are transformed into binary vectors indicating category presence or absence, is employed.

**Creating Transformers for Preprocessing:**

Transformers, which facilitate data transformations, are integral to the preprocessing steps. OneHotEncoder converts categorical variables into a one-hot encoded format, while StandardScaler standardizes numeric variables by centering on the mean and scaling to unit variance. Standardization is especially critical for algorithms sensitive to input feature scales.

**Handling Missing Values Again:**

Readdressing missing values in numeric columns is a precautionary measure. This ensures the absence of any missing values after the initial removal of the 'ID' column.

**Creating a Preprocessor:**

The Column Transformer serves as a robust tool for separately applying distinct transformations to numerical and categorical columns. This flexibility allows specific preprocessing steps tailored to each column type.

**Applying Transformations:**

The fit\_transform method is pivotal in executing specified transformations on the features, playing a crucial role in preparing the data for machine learning model training.

**Splitting the Data for Model Training and Testing:**

Dividing the data into training and testing sets is vital for evaluating model performance on unseen data. The train\_test\_split function achieves this by randomly partitioning the dataset, with the random\_state parameter ensuring result reproducibility.

# 4.Modeling

The modeling phase consists of several key steps that aim to build and fine-tune machine learning models for classification and regression tasks. The procedure is divided into the following stages:

**Data Preparation and Splitting**

The first step is to transform the raw features so that they can be used by machine learning algorithms. This critical step ensures that the data is properly structured and ready for use in the models. The train\_test\_split function divides the dataset into training and testing subsets. X\_train\_class, X\_test\_class, y\_train\_class, and y\_test\_class are defined for classification, while X\_train\_reg, X\_test\_reg, y\_train\_reg, and y\_test\_reg are defined for regression.

**Model Definition and Selection**

Among the classification models defined are RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, MLPClassifier, KNeighborsClassifier, SVC, and DecisionTreeClassifier. A selection of regression models, such as LinearRegression, is also included. Each model is accompanied by a set of hyperparameters that can be fine-tuned during the training process.

**Model Training and Hyperparameter Tuning**

The GridSearchCV method is used for hyperparameter tuning, employing a cross-validation approach to evaluate different hyperparameter combinations. This method iteratively explores the hyperparameter space to find the best settings for each model. The scoring metric used depends on the nature of the task, with accuracy used for classification models and R2 used for regression models.

**Model Evaluation and Analysis of Feature Importance**

The best-performing models are identified based on their respective scoring metrics after the hyperparameter tuning process is completed. The feature importances for these models are then extracted, revealing the most influential features in the dataset. The top features contributing to the model's predictions are displayed for models that support feature importance calculations, such as RandomForestClassifier or LogisticRegression.

So, several classification and regression algorithms were employed to predict DONR and DAMT, respectively. For classification, models such as RandomForestClassifier, GradientBoostingClassifier, Logistic Regression, MLPClassifier, KNeighborsClassifier, SVC, and Decision Tree Classifier were tuned and evaluated. Meanwhile, regression models like RandomForestRegressor, GradientBoostingRegressor, and Linear Regression were employed, each assessed and fine-tuned for performance.

A diagram of a tree classifier

Description automatically generatedA diagram of a graph

Description automatically generated with medium confidence

A graph of a number of colored squares

Description automatically generated with medium confidenceA diagram of a graph

Description automatically generated with medium confidence

A graph of a number of black squares

Description automatically generated with medium confidenceA graph of a curve

Description automatically generated

A graph of a curve

Description automatically generatedA graph of a curve

Description automatically generated

A graph of a curve

Description automatically generatedA graph of a curve

Description automatically generated

A graph of blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generatedA graph of blue dots

Description automatically generated

A graph with blue dots

Description automatically generated

# 5. Evaluation

Various metrics were used in the evaluation phase to determine how well the models addressed the goals of the nonprofit organization. Metrics like accuracy, precision, recall, and F1 score were essential for figuring out how well the models could predict donor behavior in classification tasks. While precision and recall quantify the trade-off between correctly identified positive instances and the actual positive instances in the data, accuracy shows the overall correctness of predictions. In situations where there is an uneven class distribution, the F1 score provides a balance between recall and precision that is essential. AUC-ROC, or the Area Under the ROC Curve, also shed light on how well the classifiers performed at different thresholds in differentiating between donor and non-donor instances. Together, these metrics made it possible to find models that were crucial for fundraising decisions because they balanced false positives and false negatives while also accurately predicting donor behavior.

R-squared and mean squared error (MSE) were key performance indicators in the regression field. The regression model's goodness of fit is indicated by the R-squared, which calculates the percentage of the dependent variable's variance that can be predicted from the independent variables. As for the accuracy of the model in predicting donation amounts, MSE measures the average squared difference between the actual and predicted values. The models that were found to have high R-squared values and low MSE demonstrated how well they captured donation trends and how useful they could be for precisely estimating donation amounts. This thorough evaluation of the regression and classification models addressed both technical performance and the nonprofit's financial goals by choosing models that, through their predictions, promised increased profitability.

**Model Evaluation:**

Classification Models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | AUC Score |
| GradientBoostingClassifier | 0.907 | 0.887 | 0.931 | 0.908 | 0.968 |
| **RandomForestClassifier** | 0.894 | 0.879 | 0.913 | 0.895 | 0.959 |
| **MLPClassifier** | 0.910 | 0.896 | 0.926 | 0.911 | 0.964 |
| **KNeighborsClassifier** | 0.820 | 0.758 | 0.936 | 0.838 | 0.905 |
| **DecisionTreeClassifier** | 0.851 | 0.835 | 0.872 | 0.853 | 0.877 |

|  |  |  |
| --- | --- | --- |
| **Model** | R² Score | Mean Squared Error |
| LogisticRegression | 0.170 | 44.714 |
| **SVC** | 0.450 | 29.669 |

Regression Model Evaluation:

|  |  |  |
| --- | --- | --- |
| Model | R² Score | Mean Squared Error |
| Random Forest Regression | 0.662 | 18.215 |
| **Gradient Boosting Regressor** | 0.648 | 18.955 |
| **LinearRegression** | 0.492 | 27.386 |

**Decision-Making Process:**

Best Classification Model: GradientBoostingClassifier (Accuracy: 90.42%)

Best Regression Model: Random Forest Regressor (R² Score: 66.21%)

Criteria for Selection:

Accuracy: The GradientBoostingClassifier has the highest accuracy among the classification models, indicating superior performance in correctly predicting classes.

Regression Performance: Random Forest Regressor outperforms other regression models based on the R² score, indicating a better fit to the data.

Consideration: While accuracy is a vital metric for classification, for regression tasks, the R² score is crucial to measure how well the model fits the data. Random Forest Regressor excels in this aspect.

Evaluation Conclusion:

Based on the provided metrics, the GradientBoostingClassifier stands out as the best classification model due to its high accuracy. For regression tasks, the Random Forest Regressor performs the best with a higher R² score, signifying better performance in explaining the variance in the data.

Output:

A screenshot of a computer program

Description automatically generated

**6.Deployment**

A comprehensive report was used to communicate the findings to a non-technical audience. In a new dataset (score data), the best models were used to predict donor behavior and potential donation amounts. A profit calculation was performed, taking into account the expected return on investment from targeted mailings based on the models' predictions.

**Profit Estimation:**

The script includes a process to project the expected profit from a mailing campaign using predictions generated by the top-performing models. This calculation incorporates both the forecasted donor probability (classification) and the predicted donation amounts (regression).

**Score Data Handling:**

**Data Loading:** It imports score-related data from an Excel file named 'nonprofit\_score.xlsx'.

**Column Management:** The script checks for specific columns ('id', 'donr', 'damt') in the score data and eliminates them if they exist, ensuring data consistency.

**Data Transformation:** Utilizing a preprocessor previously trained on the training data, it transforms the score data to make it compatible with the deployed models (preprocessor.transform(score\_data)).

**Model Application and Prediction Export:**

**Model Utilization:** The best classification and regression models are employed to predict outcomes based on the processed score data.

**Prediction Export:** The predictions stemming from the models (including donor likelihood and donation amounts) are incorporated into the score data.

**CSV File Export:** The finalized dataset containing the predictions is exported to a CSV file named 'model\_predictions.csv'.

In summary, this deployment script readies the score data for model predictions, employs the top-performing models, and saves the resulting predictions. These predictions can be utilized for further analysis or decision-making concerning the mailing campaign.

**7.Conclusion**

In conclusion, the analysis and modeling efforts have provided valuable insights and solutions for the nonprofit organization's goal of improving the effectiveness of their direct marketing campaigns. The exploration of the dataset revealed important patterns and relationships among demographic, financial, and donation history variables. The predictive modeling phase aimed to enhance the targeting strategy by identifying individuals likely to donate and estimating potential donation amounts.

Best Models:

The analysis identified the best models for both classification and regression tasks. The best classification model, the GradientBoostingClassifier, demonstrated an impressive accuracy of 90.42%, precision of 88.66%, recall of 93.12%, and an F1 score of 90.83%. On the regression side, the RandomForestRegressor emerged as the top performer, achieving an R² score of 65.49% and effectively predicting donation amounts.

Feature Importance:

Feature importance analysis highlighted key predictors driving the models' decisions. For example, the number of children, income, and wealth emerged as crucial factors influencing donation likelihood. These insights can guide the organization in refining their targeting strategy and tailoring campaigns to specific donor characteristics.

Expected Profit:

The evaluation of expected profits for each model revealed promising outcomes. The best classification model, GradientBoostingClassifier, is estimated to yield a profit of $6388.80, while the best regression model, RandomForestRegressor, is expected to bring in $6419.44. These projections provide a practical basis for decision-making and resource allocation.

Deployment:

To facilitate the deployment of these models, a comprehensive summary report has been generated, including descriptive statistics, best model summaries, feature importances, and model performance metrics. A graphical user interface (GUI) has been implemented for easy interpretation and sharing of the analysis results.

Recommendations and Limitations:

While the models exhibit strong performance, it's crucial to acknowledge the assumptions made during the analysis, such as handling missing data and the inherent complexity of predicting human behavior. To enhance predictive accuracy further, future research could explore advanced feature engineering techniques and more sophisticated modeling approaches.

Overall, the analysis effectively tackled the primary business question of whether accurate predictions regarding individual donation likelihood (DONR) and potential donation amounts (DAMT) could be achieved using demographic, financial, and donation history variables from the nonprofit organization's donor database. The results demonstrated that the Gradient Boosting Classifier emerged as the best model for predicting donation likelihood, achieving an accuracy of 90.4%. Additionally, the Random Forest Regressor excelled in forecasting donation amounts, boasting a strong predictive score of 65.5%. The amalgamation of these models into the best overall approach, the Gradient Boosting Classifier, proved highly promising for optimizing the organization's direct marketing campaigns to past donors. The calculated expected profits underscored the financial gains achievable through deploying these models, emphasizing their potential to significantly enhance the efficiency and cost-effectiveness of future fundraising endeavors. Despite inherent complexities in human behavior and the assumptions made in handling missing data, the outcomes offer valuable insights, empowering the nonprofit to make informed decisions and refine its strategies for improved fundraising outcomes. The detailed documentation and deployment-ready models position the organization for ongoing success in its mission.