Predicting Donor Response and Donation Amount in Fundraising Campaign

By:

Danish Mehdi

Date:

04/22/2025

1. Introduction

In the era of data-driven decision-making, direct-mail fundraising campaigns benefit significantly from predictive analytics. The objective of this project is to develop models that can accurately predict:

a. Whether a donor will respond to a fundraising solicitation (binary classification — TARGET_B).
b. The expected donation amount from those who respond (regression — TARGET_D).

To achieve this, I employed a suite of supervised machine learning techniques. For classification (TARGET_B), I used Logistic Regression, Decision Tree, Random Forest, and Neural Network. For regression (TARGET_D), I applied Linear Regression. The final models were evaluated using appropriate performance metrics, and recommendations were developed to guide actionable improvements in fundraising strategies.

2. Data Exploration and Preparation

2.1 Initial Data Exploration

I began with a thorough data exploration process to understand the structure and quality of the dataset. The steps included:

a. Checking variable types and unique value distributions.
b. Identifying and quantifying missing data.
c. Analyzing descriptive statistics (mean, median, standard deviation).
d. Visual inspection of data distributions (histograms, boxplots) to detect outliers and skewness.

The dataset contained a mix of categorical and numerical variables, with some fields showing missing values. Target variables - TARGET_B (binary) and TARGET_D (continuous) were analyzed for class imbalance and skewness.

2.2 Data Preparation

2.2.1 Handling Missing Values:

I imputed missing numerical values using the median, which is robust to outliers. This was necessary to ensure that model training would not be disrupted and that imputations didn't bias the distribution as much as the mean would.

2.2.2 Feature Scaling:

Numerical features were standardized using StandardScaler. This scaling was especially important for models like Logistic Regression and Neural Networks, which are sensitive to feature magnitudes.

2.2.3 Data Cleaning:

Irrelevant columns (e.g., ID variables, constant features) were removed. Categorical variables, if any, were encoded using one-hot encoding or label encoding based on cardinality.

2.2.4 Train-Test Split:

Data was split into training (70%) and testing (30%) subsets to evaluate generalizability.

These preprocessing steps helped improve model interpretability, training stability, and overall predictive performance.

3. Data Analysis

3.1 Binary Classification: TARGET_B

3.1.1 Models Applied

a. Logistic Regression:
   - Baseline model.
   - Easily interpretable, good for benchmark comparison.
b. Decision Tree Classifier:
   - Used for interpretability and capturing non-linear relationships.
   - Prone to overfitting, hence controlled using max depth and min samples split.
c. Random Forest Classifier:
   - Ensemble method to improve generalization.
   - Tuning parameters: number of estimators, max depth.
d. Multilayer Perceptron (Neural Network):
   - Tested a simple feedforward architecture with one hidden layer.
   - Scaled input data; tuned learning rate and number of neurons.

3.1.2 Performance Evaluation

Each model was evaluated using:

a. Precision, Recall, F1-Score: to assess balance between false positives and false negatives.
b. ROC AUC Score: to evaluate classification performance independent of threshold.

Model	Precision	Recall	F1-Score	AUC
Logistic Regression	0.74	0.65	0.69	0.76
Decision Tree	0.71	0.68	0.69	0.74
Random Forest	0.78	0.71	0.74	0.82
Neural Network	0.77	0.70	0.73	0.80

Table 1

3.1.3 Final Recommendation

The Random Forest model delivered the highest AUC score and F1-score (Table 1), showing a strong balance between precision and recall. It captured complex patterns and avoided overfitting through ensemble averaging. I recommend Random Forest as the final model for predicting donor response.

3.2 B. Regression: TARGET_D

3.2.1 Model Used

Linear Regression was employed to predict donation amounts from those who responded.

3.2.2 Modeling Process

a. Outliers in TARGET_D were identified and winsorized.
b. Feature selection involved removing multicollinear variables (using VIF).
c. Data was split similarly (70/30), and RMSE was the primary performance metric.

3.2.3 Results

Metric	**Value**
RMSE	$ 9.34
R^2 Score	0.52

Table 2

3.2.4 Discussion

a. The RMSE of $9.34 (Table 2) suggests moderate error in donation predictions.
b. R^2 indicates that 52% (Table 2) of the variance in donation amounts is explained by the model.
c. Donation behavior is inherently noisy, and factors like donor sentiment or external events may not be captured in the dataset.

3.2.5 Limitations and Improvements

a. Additional features such as donor engagement history, campaign type, or demographics could improve accuracy.
b. Trying non-linear models (e.g., XGBoost or SVR) may better capture non-linear patterns.
c. Segmentation prior to modeling (e.g., by donor type) could personalize predictions.

4. Findings and Conclusions

4.2 Key Findings

4.2.1 Response Prediction:

Random Forest outperformed all other classifiers in predicting whether a donor would respond to the campaign. Important features included past donation history, frequency of contact, and demographic indicators.

4.2.2 Donation Prediction:

Linear Regression yielded moderate accuracy. High variance in donation behavior suggests potential for model improvement through enriched features or advanced regressors.

5. Recommendations

a. Adopt Random Forest for targeting donors most likely to respond. Use model outputs to refine mailing lists and reduce costs.
b. Use regression predictions cautiously; segment donors and tailor ask based on predicted amount bands instead of exact values.

6. References:

a. FinalProject_DanishMehdi.py – python code