Overview of the Problem To determine which potential donors to solicit for maximising the profitability of the charity's marketing campaign.

Business Understanding
Problem Statement - A marketing Strategy to choose potential donors such that Gross profit (Donations - Marketing cost) is maximum.

Data Understanding
data/train.csv

> 2 Independant variable - Amount (continuous numerical) and Responded (binary classification)
> 3 primary sets of 478 dependant variable - Characteristics of the customers neighborhood~280, Promotion and giving history file~100,Other features~100

Data Preparation
Mean Imputation - If less than 30% of a variables values were missing (and at random) they were replaced with the coloums average value.
Zero Filling - where appropriate, missing values were filled with 0.

Feature Selection
Principal Component Analysis (PCA) - To reduce the dataset's dimensionality while retaining most of the information.
Dropped irrelevant variables - whose description meant no relevance.
Dropped redundant variables - but retained its summary rows.
Dropped highly correlated variable - to reduce multicollinearity.

Modelling
Approach -

1. First we build a classifier model to predict the likelihood of a user (potential donor) responding positively (i.e., making a donation) (Pi)
2. Given that the user donates, we predict the amount they will donate (Ai)
3. For each user, you calculate the expected donation value E(Di) by multiplying the probability of donation with the predicted donation amount E(Di) = Pi*Ai
4. If marketing cost to call the user is Mi, we choose to call those users whose expected outcome is positive.ie, E(Di) - Mi > 0

Process -
The classification model used is a Random Forest Model which was trained over 90% of the dataset, hyperparameters tuned using 5 fold Cross Validation. Random Forest was chosen due to the large number of features including categorical variables and the absence of evident and explainable patterns or distribution.

The second model is a SImple Vector Regressor which predicts the amount of donation for the respondents. SVM was chosen as this amount need not be pinpoint accurate but predicts in an acceptable range. And also for the fact that errors within the epsilon-insensitive margin do not contribute to the loss function and hence provide a control on acceptable error range (Epsilon and C).

Evaluation

Model Performance and Results

The Classification model was optimised for maximum recall over precision since we need to minimise the no of False negatives which are opportunity costs. This will in turn maximise the coverage of calling and hence more donations. The classification threshold was adjusted using F2 score, giving higher importance to recall.

Current status - Precision - 100%, Recall - 66%

Whereas the regression model trained on L1 errors stands and 23% MAPE.

The above strategy will cover 7% of the donor base in which 98% is expected to donate compared to 5% conversion if we had called all of the users.

Conclusion

This approach provides a framework for data-driven decision-making where each potential contact is evaluated in terms of the Expected Utility versus the Incurred Cost. The approach ensures that the resources for outreach are allocated in the most efficient manner, prioritising those individuals who are most likely to yield the best return on investment.

Next Steps: The model can be further improved by more detailed EDA and Feature engineering