Maximizing Donations through Targeted Solicitation

PREMASHISH MUKHERJEE, George Mason University, G01390485, pmukher@gmu.edu, USA
DEEP VORA, George Mason University, G01388910, dvora@gmu.edu, USA
SAURAV SINGH, George Mason University, G01390520, ssingh89@gmu.edu, USA

Additional Key Words and Phrases: cost sensitive learning, kdd data set, random forest

Reference:
Premashish Mukherjee, Deep Vora, and Saurav Singh. 2023. Maximizing Donations through Targeted Solicitation . 5 pages.

Authors' addresses: Premashish Mukherjee, George Mason University, G01390485, pmukher@gmu.edu, Fairfax, USA, pmukher@gmu.edu; Deep Vora, George Mason University, G01388910, dvora@gmu.edu, Fairfax, USA, dvora@gmu.edu; Saurav Singh, George Mason University, G01390520, ssingh89@gmu.edu, Fairfax, USA, ssingh89@gmu.edu.

1 PROBLEM STATEMENT

This project is focused on developing machine learning models that accurately predict donors and donation amounts, with the ultimate goal of maximizing donations solicited through marketing emails for retired veterans.

The main challenge in our problem definition is due to the fact that our dataset has imbalanced class distribution. This imbalance can lead to a biased model towards the majority class. Our main goal is to correctly identify the potential donors which is the minority class because they directly impact the donation amount. Hence, we will use Cost Sensitive Learning for this Imbalanced Classification.

Cost Sensitive learning for Imbalanced Classification involves assignment of different cost to different classification errors and then incorporating those cost when classification is done. We will analyze the different cost sensitive learning techniques and compare their evaluation metrics.Finally ,based on our analysis of the different evaluation metric we will select the best performing model.[1]

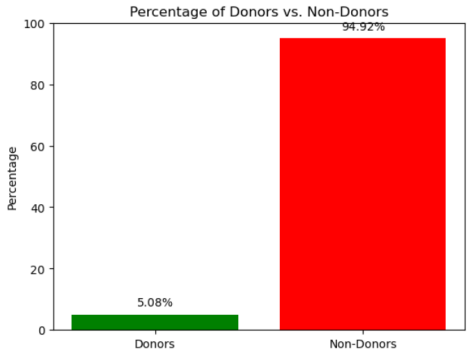


Fig. 1. Distribution of Donors

2 METHODOLOGY

Our approach to solving this problem of imbalance class distribution of the data set can be categorized into three major components. We had 2 main files to read in here, one for learning (cup98lrn.txt) and one for validation

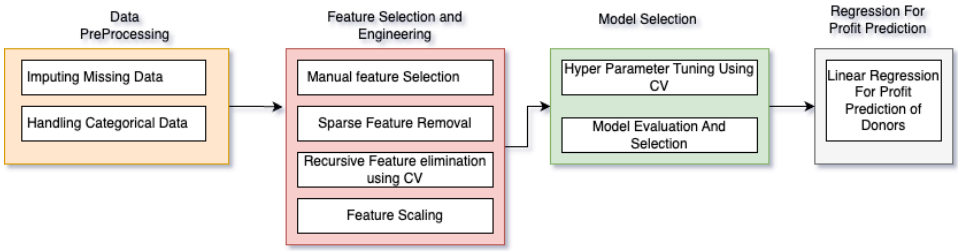


Fig. 2. Methodology Flow Diagram

(cup98val.txt). These files are too large and have huge amount of features and hence various data prepossessing and feature selection techniques was used as mentioned below :

2.1 Data Preparation

2.1.1 Imputing Missing Data. Data had lot's of missing values which need to be handled hence we imputed the missing values having object data type such as categorical values etc. with most frequent values where as for other data type we replaced them with the median.

2.1.2 Handling Ordinal and Nominal Values. Data include both ordinal and nominal type of categorical values which need to be handled, hence we converted them to one hot vectors which resulted in further increase of the number of features. These features also resulted in further adding sparsity in the data.

2.2 Feature Selection

2.2.1 Manual feature selection. In feature selection we removed some of the features by manual dropping, This includes Census data which conveyed information about the neighbourhood but not about the individual itself, training on these features results in model which gets highly biased towards neighbourhood hence deviating from individual assessment.

2.2.2 Removal of Sparse features. To reduce the natural sparsity of the original data as well as sparsity introduced due to one hot vectors we removed the features which had more than 99.9 sparsity.

2.2.3 Recursive feature elimination with cross-validation. Model is first trained on some initial set of features selected through above mentioned steps, importance of each such features are then obtained through feature importance attribute and coefficient score. These scores then help us in pruning the least important features, this method occurs recursively and in the end we find the desired number of final features.

2.2.4 Standardization of features and Removal of Highly co related feature. Data had vast of Range of values hence we tried training our model with both min-max standardisation as well normal standardisation. We also removed some highly correlated features for further reducing the high dimensions.

2.3 Model Selection and Classification of Imbalance data

After our data was prepared we needed to train our model to predict the donors and the non-donors. Classification was done on unbalanced data and balanced data which was obtained through variety of techniques such random down sampling, automatic balancing by models as well manual weight assignment using cross-validation. We also did cross validation for Hyper parameter tuning of our models. We majorly focused on three main models logistic regression, Random forest and ensemble method which combined previous two models.

2.4 Regression

Lastly we passed our data through a regression model for predicting the amount of donations for addressing the second part of our problem. These regression metric helped us in deciding donation amounts which were crucial in deciding whom we need to mail. For regression we majorly on Linear regression.

3 EXPERIMENT AND ANALYSIS

This section includes the results, and analysis on KDD98 data using our methodology. We first trained our model using the 80% of our train data(cup98lrn.txt), the rest 20 % we used as our validation-set. We provided experimental results for both the validation set as well as for the test-set(cup98Val.txt), we authenticated the test set result using valtargt.txt for different metrics and for calculating the actual profit of our model.

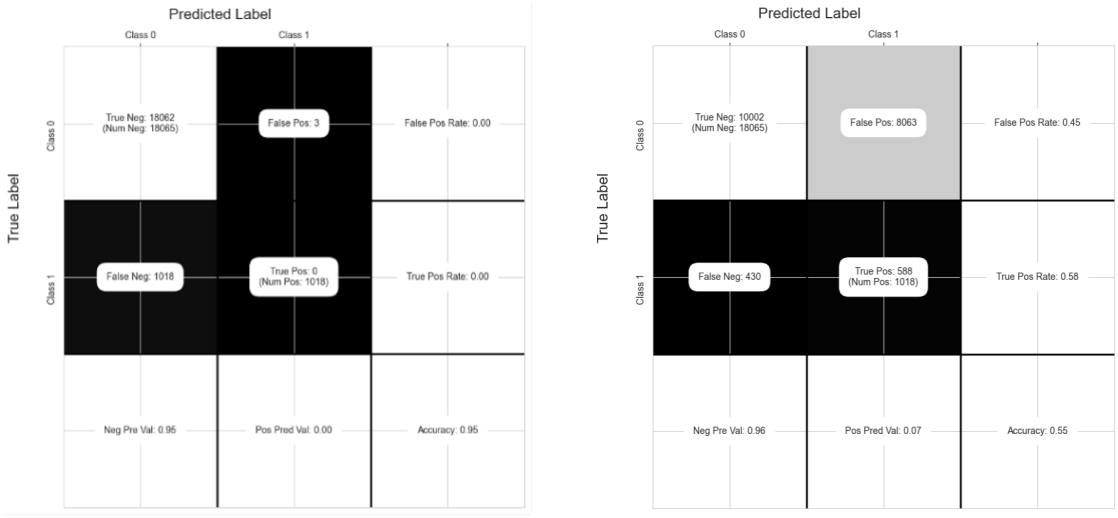
Train Set	Validation Set	Model	Balancing	Accuracy	F1-Score	Precision	Recall	Profit
76329	19083	Logistic Regression	Not Balanced	94.64	0	0	0	-2.04
76329	19083	Logistic Regression	Balanced by model	62.86	12.3	7.04	48.9	1618.479
76329	19083	Logistic Regression	Balanced by Random-Downsampling	55.49	12.16	6.7	57.7	2445.31

Table 1. Metric Evaluation on Validation set for different Sampling

3.1 Balancing/Sampling

As we can see, the accuracy on Unbalanced data set is high, but the profit is negative. This is because the data is highly imbalanced and our model is not doing a good job in predicting the minority class. This is evident from the F1 score which is zero.

Our model seem to perform better when we assign class weights according to the distribution of the 2 classes. There is significant improvement in the F1 score and hence the profit on the validation set. We can also balance the data set by downsampling of the negative examples. Our model seem to perform even better in this case. The Confusion matrices below summarizes our interpretations



(a) Confusion Matrix For Unbalanced DataSet

(b) Confusion Matrix for Balanced Dataset

Fig. 3. Confusion matrix comparison

3.2 Model Selection

Validation Set	Model	Accuracy	F1-Score	Precision	Recall	Profit
19083	Logistic Regression	55.49	12.16	6.7	57.7	1674.59
19083	Random Forest	56.35	12.75	7.14	59.8	2445.31
19803	Ensemble of LR and RF	53.95	12	6.67	58.84	1540.37

Table 2. Metric Evaluation on Validation set for different Models

Logistic Regression and Random Forest seems to have comparable results in terms of the evaluation metrics. So, we ran both the models on the test set and found out that Random Forest yields higher amount of donation. Also it is evident from the AUC plot that Random Forest is better at classifying the instances.

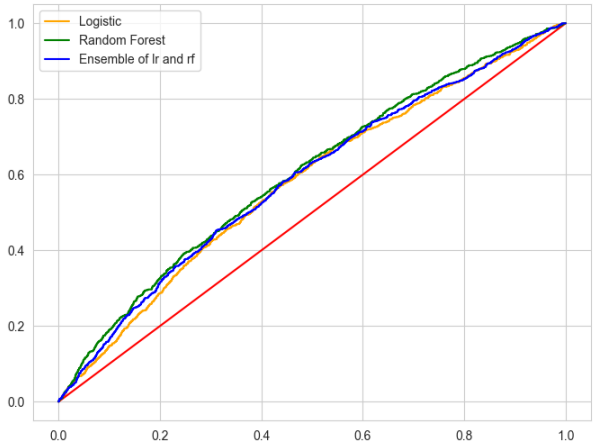


Fig. 4. ROC curve across 3 Models

3.3 Test Data Result

The test data comprises of 96376 instances. We ran the RandomForest classifier on test dataset and predicted the donors. Finally, we trained the Linear Regression model on our training dataset and then predicted the donation amount of the potential donors.

Test Set	Model	Accuracy	F1-Score	Precision	Recall	Profit
96367	Random Forest	55.57	12.05	6.7	60	8368.55

Table 3. Metric Evaluation on Test set

4 CONCLUSION

Such type of problems in which we have imbalance class distributions, often require a thorough assessment of different Cost Sensitive Learning techniques. In such type of problems accuracy is not the goal. So, we need to analyze the F1 score which takes into account the precision and recall.

Through our analysis, we were able to improve the recall, however increase in precision is relatively low, due to the high number of false positive cases. This is still a challenge for us and a hindrance in profit optimization and hence leaves room for future work.

Based on our analysis of different techniques, RandomForest classifier with resampling is giving the best results.

REFERENCES

[1] Jason Brownlee. 2018. Cost-Sensitive Learning for Imbalanced Classification. (2018). https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/.