COST-SENSITIVE LEARNING

Machine Learning on Imbalanced Datasets

KRUTHAY KUMAR REDDY, DONAPATI
MS CS Student, George Mason University, kdonapat@gmu.edu

This report describes our effort to develop a model that could maximize the profit generated from the 1997 solicited donation campaign data.[1] Several feature selections and data cleansing techniques such as under-sampling are used to reduce the imbalanced nature of the data and then model selection techniques such as cross-validation and voting classifiers are used to predict the possibility of a donation and the amount donated. Based on the predictions of our model, soliciting the predicted donors can generate a profit of $10869, a little better than the profit gained by mailing every person.

CCS CONCEPTS • Computing methodologies ~ Machine learning ~ Machine learning approaches

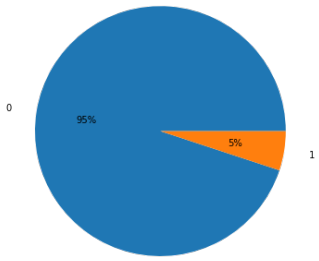
Additional Keywords and Phrases: Imbalanced Data Set, Voting Classifier, Rank based Predictions, False Negatives

ACM Reference Format:

Kruthay Kumar Reddy, Donapati. 2022. Cost-Sensitive Learning, Machine Learning on Imbalanced Datasets, May 13, 2022, Fairfax, Virginia, 5 pages.

1 Introduction

A solicitation was mailed to people expecting a donation from them. However, only 5% of the people donated. It costs 0.68 cents to solicit. The average donation received using the "Mail to All" strategy is 0.79 cents. The net assistance received after removing the costs is $10788 for the training set. The maximum profit that can be gained by only soliciting the donors is $75668.7.



Difference between the percentages of donors and non-donors

Our motivation is to develop a machine-learning model to predict the donors and the amount they have donated. We want to verify that it's possible to create a model that can improve the profits by reducing the false negatives and considering the model's reliability in predicting a future unknown dataset.

The data provided is imbalanced, with many outliers in many features. There is little to no correlation between characteristics and target variables, making it hard to train the model. We studied each feature individually using the provided feature description with the dataset and performed the statistical analysis. We then used our knowledge to select a sample of features for training the model.

We aim to reduce the apparent nondonors so that gain could be increased. If we find 8% of obvious negatives, we can improve the yield by 40%. We have used classification models to remove the evident non-donors from the data based on prediction probabilities. We repeated the above step multiple times, and the top-ranked apparent known donors were removed to under-sample that data. We then used the Voting classifier [2] on a few models to predict the donor and ranked the outputs based on the probability. The probabilities are chosen to reduce the false negatives on the training data.

Due to the imbalanced nature of the data, traditional evaluation metrics don't provide routine evaluation. We used the ratio of false negatives to true positives as our evaluation metric. Low scores indicate that the apparent nondonors are neglected from mailing, thereby increasing profits.

If the solicitations are only sent to the predicted donors, the profit will be $10869. As expected, the model performed poorly on traditional evaluation metrics such as accuracy, confusion matrix, and f1 scores. However, it performed better on our metric with 0.0014 scores on false negatives/ true positives.

2 Method

2.1 Dataset

The training dataset consists of 95412 examples with 479 features and 2 labels. According to the feature description, the dataset is collected from two sources, Metromail and Polk. The label TARGET_B indicates if a donor has donated, while TARGET_D provides the donated amount.

Various numerical and categorical features are provided in the dataset describing the donor's age, income, wealth, title, homeownership, family information, response to other emails, interests, and neighborhood information. A few other fields describe the recency, frequency, and amount of past donations in response to various mailings. Almost all features offer very little to the target variables. Most of the features have invalid data of more than 50%.

2.2 Feature selection

The dataset has many outliers; for example, the Gender feature has 8 unique values with no information on how to transcribe a few values. It took much effort in data cleansing for this data set. We grouped certain features based on our observation and dealt with those features as required; for example, there are 14 features with exactly 52854 values missing, and we found that this group of features offers no correlation with the target variable, so we neglected those features.

2.2.1 Selected Features.

After careful consideration of each feature based on correlation and feature description, we have selected the following features to train the model: 'ZIP', 'INCOME', 'HIT', 'NUMPROM', 'RAMNTALL', 'NGIFTALL' , 'LASTGIFT', 'LAST DATE', 'HOMEOWNER,' 'MAXRDATE,' 'MAXRAMNT,' 'RAMNT_14', 'HV2', 'AVGGIFT,' 'RP2', 'MAJOR,' 'IC4', 'AGE,' 'MINRAMNT,' 'RFA_2F', 'RFA_2R', 'RFA_2', 'MDMAUD_R,' 'MDMAUD_F,' 'MDMAUD_A.'

2.3 The Evaluation Pipeline.

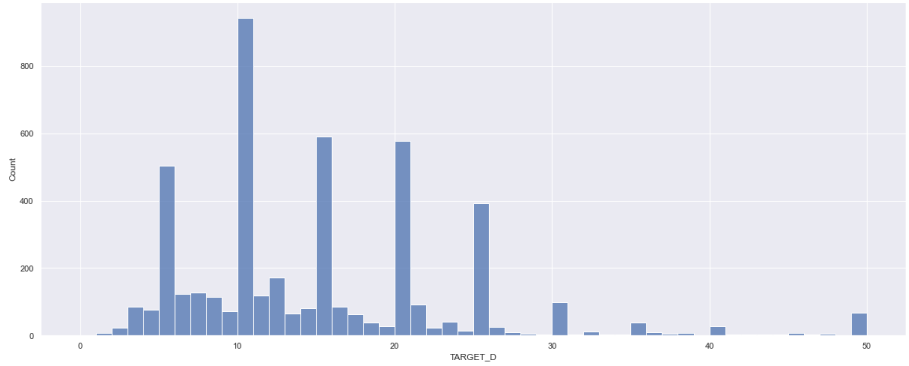
One of the ways to deal with an imbalanced dataset is to train the model with more data. Hence, we split the training dataset into a 9:1 test train split to maximize the available data. We've tried to use many under-sampling techniques such

as Random Under Sampler [3], Edited Nearest Neighbor [2], and oversampling techniques such as SMOTE[3], but they have not improved our model.

2.4 Methodology

2.4.1 Preprocessing

We've used our statistical analysis to fill in the values for the features with null values. We've filled the null values based on our observations, for example, 'RFA 2R' has letters A to G, with each value providing the range of the donation amount, and these values are replaced with the lowest range value. All the feature flags with letters are mapped to binary values. Values in TARGET_D are grouped as 0, 1 and 2, where 0 represents donations less than $5, 1 illustrates donations in between [5, 25) and 2 means contributions more significant than 25



Density distribution of the Donated amount

2.4.2 Model Selection

We've used the Adaboost classifier [2] to remove the apparent negatives. We removed the instances predicted to be negative with a probability of 0.98 and under-sampled the data; we repeated the above until the non-donor samples were reduced to 30% or the maximum iterations were reached. We have used 10-fold cross-validation on various models with different parameters to determine the better-performing models. Most of the models were poor, while others were marginally better. We've then selected these models to use in the voting classifier and predicted the donors based on the probability threshold. We chose the entry as 0.82 so that the false negatives to true positives ratio is minimized. Linear regression [2] is applied to the instances with predicted donors. However, as the classification accuracy is inferior, the model didn't perform as expected.

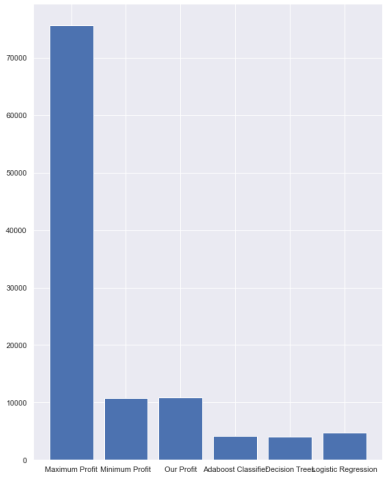
3 Result

3.1 Table

Model	Accuracy	Profit	F1 score	Chosen Metric	Accuracy	Profit	F1 score	Chosen Metric
	Training Data				Test Data			
Logistic Regression	57.52	$4985.13	0.11	0.80	56.25	$4679.03	0.09	1.312
Decision Trees	64.16	$18776.5	0.16	0.475	61.19	$4045.6	0.09	1.632
Adaboost Classifier	61.84	$8430	0.13	0.72	59.77	$4097.2	0.08	1.55
Voting Classifier with under sampling	47.12	$12892.43	0.04	0.0002	45.74	$10869.1	0.04	0.0014

Table with Results of Training and Test Data

The results could be more suitable as the profits generated are far from the maximum achievable profit. However, our model performed better than plain classification models. We used a linear regression model to predict the level of donors based on the predictions from our voting classifier, and it produced an accuracy of 12%. However, we've trained the linear regression model on positive instances in the training data and tested it against the test data, producing an accuracy of 73%. The reason for the poor performance has to do with the poor accuracy of the voting classifier. Due to the high complexity of our model, it may not be reliable to deal with a future dataset.



Comparison of profits on various models

4 Conclusion

Imbalanced data is tough to deal with, particularly if the dataset is skewed. Many techniques are available to deal with the imbalanced nature, and different evaluation metrics must be used. Based on the requirement, one can choose the evaluation metric which suits the objective.

We would have spent more time on feature selection and preprocessing. We would use the neighborhood data to create clusters and fill in AGE and INCOME values based on that cluster information. We could have tried finding patterns in donors' donation frequency.

If we consider human emotions, we can probably use DOB to predict a donor's donation pattern, which would have been interesting. However, the month column in DOB is inappropriate, with more than 50% outliers.

REFERENCES

[1] Ismail Parsa, Epsilon, 50 Cambridge Street, KDD Cup 1998 Data. https://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html</bib>
[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay; Scikit-learn: Machine Learning in Python. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
[3] Guillaume Lema, Fernando Nogueira, Christos K. Arida, Imbalanced-learn: A Python Toolbox to tackle the Curse of Imbalanced Journal of Machine Learning Research, 2017. https://imbalanced-learn.org/stable/about.html