Predicting Long-term Deposit Subscription From Bank Telemarketing

Matthew Edwards *
Dept. of Statistical Sciences
University of Toronto
1000551333
mr.edwards@utoronto.ca

Amir Masud Zare Bidaki
Dept. of Computer Science
University of Toronto
1003828012
amirmasud.zare@mail.utoronto.ca

December 2020

Abstract

Using data on telemarketing campaigns from a Portuguese Bank, we predict whether a customer will agree to subscribe to a long-term subscription. Using over 40000 observations of 20 different features, we fit several baseline models. Using the fact that our features split evenly into groups, we also fit an ensemble method tailored to the dense but unevenly distributed numeric variables, and the sparse categorical ones. We find an ensemble method using a decision tree on the first group and a neural network on the second gives near state-of-the-art results without overfitting.

Introduction

Marketing is often a gamble for businesses, because they cannot be sure the customer will be receptive to their ideas. This uncertainty makes it difficult to attract new business. Although each company has its own marketing strategy, many have found that customers value direct communication with other people. In other words, human connection plays an important role in customer service. Consequently, virtually all companies use some form of direct marketing (e.g. email. telephone, social media) to try and connect to their market segment.

Perhaps the largest problem with this approach is that it is very inefficient; businesses can waste hours on customers who simply are not interested in transacting with them. Thus companies would save a lot of time (and by extension, money) if they could target only the customers they knew would be willing to buy the product. So how do they know which customers are more receptive? What makes the ideal customer? In machine learning terms, this is a simple *classification task*: Given the characteristics of the customer and the business, we want to predict whether the customer is likely to buy the product or service.

*For all code related to this paper, see the github repository: https://github.com/12mre1/csc2515F-final-project

In this paper, we use data on telemarketing phone calls from a Portuguese bank to try and predict whether customers decide to subscribe to a long-term subscription option offered during the call. Using 20 features based on client information, data from previous contact, and socioeconomic indicators, we use an ensemble method where, instead of just fitting models on the basis of overall performance, we fit models to different subsets of the features based on feature type.

Related Work

Much work has been done on predicting success or failure based on marketing strategies. Moro et al. (2014) use a similar dataset to fit several different models to predict success or failure [1]. They find that, among several different classes of models including linear classification, support vector machines, and others, neural networks give the best performance based on classification accuracy, achieving just over 90 %. However, when Laureano et al. use a slightly wider dataset to model the same task, they find instead that SVMs provide the better performance, albeit on a smaller training set, used because of computational constraints [2].

Consensus on which model works best has not yet been reached. This is because acquiring labelled bank data is often very costly. Consequently there is a shortage of large labelled training data. This means that algorithms with more flexibility ideal for big data (ie neural networks) are unable to differentiate themselves from classical approaches. It also implies that performance can vary widely depending on features, and evaluation. For example, Hassan et al. find that simple Logistic Regression works best on a dataset of slightly different features derived from ours [4]. Recognising that the most obvious way to improve performance is to use more training data, several studies have attempted to augment existing datasets using semi-supervised and simulated approaches [8].

Our approach does not involve changing the dataset. Instead we fit different types of models on different groups of features. Recent work has revealed that certain types of features lead to increased performance on certain model types [3]. For example, neural networks are better equipped to handle large, sparse features, while tree-based methods are excellent at classification when features have uneven distributions. With this in mind, our ensemble method uses a decision-tree classifier for uneven numerical features, and a neural network for the sparse categorical ones. Such combined methods have a long history in tasks whose goal is purely predictive[7].

Data

The dataset comes from the Machine Learning Group at the University of California Irvine [1]. It is a condensed version of the dataset used by Moro et al (2014) [1]. We use a total of 41188 observations, of which 5000 are set aside for testing. In total, we have 20 different features, on a wide variety of topics:

i) Client Data: *age, job, marital status, education, default on loans (0-1), default on housing payment (0-1).*

[1] You can download the data directly by following this link: `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`

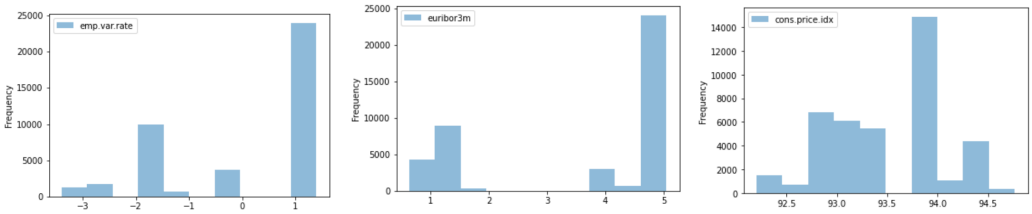

Figure 1: Histograms of Unemployment Variation, Euribor, and CPI.

ii) Marketing Data: *Previous contact (0-1), number of prev. contacts, days since last contact, length of previous contact, previous outcome (0-1).*

iii) Socioeconomic Data: *Employment variation rate, Consumer Price Index, Consumer Confidence Index, Euribor Rate, Number of people in household employed.*

iv) Seasonal Data: *Day of Week, Month, Time of Day.*

The response variable is binary, indicating whether the customer agreed to subscribe to the long-term deposit after the call. Since, after inspection, the datset do not appear to be sorted in any discernable way, we split off the test data simply by removing the first 5000 observations. Among the training data, we see that y is distributed with 4640 positive and 36548 positive cases. These percentages work out to roughly 13% and 87%. This means that, if we want to use accuracy as a measure of performance, the floor should be at least 87%, since this can be achieved with a naive classifier predicting no regardless of features. Fortunately, there are no missing values in any of our features, so no observations are imputed or dropped.

On further examination of the dataset, we see that the features split nicely into two types: categorical (11) and numeric (9). This means we have a fairly balanced number of features on which to fit our two different classes of model. Moreover, among the numeric features, we find the distributions are very uneven (see figure 1). This makes them ideal for classification using tree-based algorithms [5].

Models

To improve performance, we use one-hot encoding on our categorical variables. We also use min-max scaling on our numeric features (since many are integer-valued only, standardization would not be as effective as usual). This results in the original 20 features being turned into 62 (9 numeric and 53 categorical). To put the performance of the ensemble approach into context with the other state-of-the-art approaches, we also fit a series of baseline models based on previous work:

1. Logistic Regression using all features (this is the approach from [4])
2. Decision Tree Classifier with just numerical features
3. Neural Net with just categorical features
4. Decision Tree with all features
5. Neural Net with all features (this is the approach from [1])

All models are evaluated with 10-fold cross-validation. We use as performance metrics both accuracy, and area under the Receiver-Operator Characteristic Curve (AUC score). Hyperparameters are chosen by grid-search, accounting for various regularization coefficients, solvers, numbers of hidden layers, sizes of hidden layers, loss functions, tree depths, split criteria and regularization techniques.

Results

Table 1: **Performance Metrics and Hyperparameter Settings for Various Models**

Model (features)	Hyperparameters	AUC	Acc.
Logistic Reg (all)	L1 Reg.($\alpha = 10$), SAGA solver, 300 iters	0.76	0.90
DT (numerical)	Depth(5), Features (4), Gini Crit.	0.78	0.90
NN(categorical)	1 Hid. L, 8 neurons, N(0,1) W init.	0.76	0.89
DT (all)	Depth(5), Features (4), Gini Crit.	0.71	0.90
NN (all)	1 Hid. L, 8 neurons, N(0,1) W init.	0.77	0.90
DT(num) + NN(cat)	Same as above, DT weight(9/20)	0.8	0.92

Looking at Table 1, we see some interesting results. The logistic regression baseline performs just as well as either the Decision Tree with only numeric features, or the NN with just categorical features. When we combine the features into one set, the neural network performs slightly better in terms of AUC score, but does equally well in terms of accuracy. Finally our ensemble model is the best amoung them. It achieves an AUC score of 0.8, which is on par with the state of the art [2]. It also does better than all the baselines in terms of accuracy, however other studies have reported an accuracy of above 0.95 in some cases [6]. Overall, the ensemble approach has performed well relative to both the baselines, and to other studies.

Discussion and Future Work

Although these results are promising, there is still room for improvement. Compared to other state of the art approaches, our dataset is fairly thin. Some work in feature engineering could increase performance. For instance Moro et al. start from a dataset of over 110 features, whereas we use only 20. Another area of improvement would be to refine the techniques of the decision tree classifier. Bagging or boosting may add additional performance, as would the use of a Random Forest Classifier. It is also interesting to note that the performance of the neural net using only the categorical features was among the worst, suggesting that perhaps the numeric features are much more informative in predicting success or failure. Given that the numeric features are mostly socioeconomic, this suggests that whether a person is willing to buy a product is largely driven by environmental factors (ie this may be mostly a spatial-temporal problem). The choice may less to do with the individual situation.

One last (and perhaps the most obvious) way to improve performance would be to collect more training data. However, despite recent attempts by many large companies to collect big data, we must be careful about the conclusions we draw. There is still a proportion of

the population, particularly those who are highly introverted, who are unlikely to respond to the types of direct communication used to collect samples like the data we've used. Thus both selection and sampling bias affect the generality of our conclusions.

References

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.

[3] Li P. et al. Combining Decision Trees and Neural Networks for Learning-to-Rank in Personal Search. 25TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2019)

[4] D. Hassan, A. Rodan, M. Salem and M. Mohammad, "Comparative Study of using Data Mining Techniques for Bank Telemarketing Data," 2019 Sixth HCT Information Technology Trends (ITT), Ras Al Khaimah, United Arab Emirates, 2019, pp. 177-181

[5] Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010.Adapting boosting for information retrieval measures. Information Retrieval13,3 (2010), 254–270

[6] Bhaskar Mitra and Nick Craswell. 2017. Neural Models for Information Retrieval. arXiv preprint arXiv:1705.01509(2017).

[7] Yu, J. M., Cho, S. B. (2016). Prediction of bank telemarketing with co-training of mixture-of-experts and MLP. In K. Ikeda, M. Lee, A. Hirose, S. Ozawa, K. Doya, D. Liu (Eds.), Neural Information Processing - 23rd International Conference, ICONIP 2016, Proceedings (pp. 52-59).

[8] Lucas Ruff et al(2020). Deep Semi-Supervised Anomaly Detection. International Conference on Learning Representations (2020).