A Comparison of Logistic Regression and Random Forest for a Bank Marketing Data Set

SI: 210002017

1. DESCRIPTION OF THE PROBLEM

We solve a binary classification problem by predicting if a customer will subscribe to a term deposit using information only available before a marketing call is made. Our main aim to check the suitability of models with embedded feature selection — a regularised logistic classifier and a random forest model — and compare the results to previous work which used manual feature engineering [1] [2].

2. DATASET AND EXPLORATORY ANALYSIS

- The Dataset consists of 41,188 observations and provides data of individual characteristics of consumers as well as some general economic values. The dependent variable is whether a consumer will subscribe to a term deposit—either yes or no. The data is highly imbalanced with only 4640 people subscribing yes for 36540 no's.
- There were no missing values in the dataset.
- We decided not to keep two variables from the original data set. The first was 'duration' which was the duration of the call of a successful sale for a term deposit. As it cannot be known beforehand we felt it was superfluous for prediction purposes. The second was the binary variable 'default' which was an almost constant variable.
- There were no missing observations although there were a total of 12000 observations which had 'unknown' in one of their categorical predictors. We decided to use these as a class label. An additional option could have been to use mode imputation but we thought our's was a reasonable solution considering it is common in marketing for data on individual consumers to be highly siloed.
- We then created dummy variables: one for each factor of each categorical variable. All the predictors were them normalised to have mean zero and standard deviation. Although normalisation is not required for logistic regression (or random forest) and there is some debate about normalising categorical variables [3], it is required for some processes' such as the L1 norm. Hence we totalled 59 predictive variables in addition to the dependent variable.

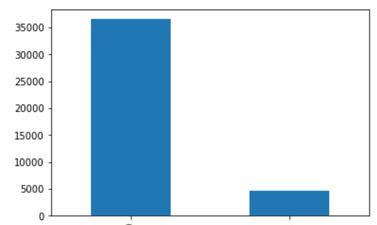

Figure 2: Distribution of the dependent variable

Figure 1. Correlation of the numerical predictors

- Moreover many of the predictors suffered from Collinearity as can be seen in correlation heat map is shown on Figure 1. Moreover many of the categorical variables suffered from this as well. For example p-outcome — whether a call was a success or failure—is highly correlated 'previous' — the number of contacts performed before this campaign. We will try to alleviate these through methods below.

3. THE TWO MODELS

LOGISTIC REGRESSION

- Logistic Regression is a supervised learning technique which uses a logistic function to model a binary dependent variable on a set of independent variables.
- It predicts the posterior probabilities of the target class as a linear function of the predictors. A suitable decision boundary can then be used to classify the binary variables.
- It can be extended for case of multi class classification.
- Similarly to linear regression, regularisation techniques such as L1 or L2 can be applied to prevent overfitting.

Advantages

- Simple and easy to infer as it is computationally efficient and provides coefficient size as well as direction.
- Advantageous as it also provides probabilities.

Disadvantages

- The decision boundary is linear and hence cannot solve non-linear problems; linearly separable data is rarely found in real world scenarios.
- Does not perform well with overlapped classes or co-linearity between predictor variables

RANDOM FOREST

- Random Forest constructs multitude decision trees—which each provide their own predictions—and minimises error through a decision rule. Using the CART algorithm, each tree randomly samples form the observations and the predictors. For classification a prediction decision is achieved through a majority vote.

Advantages:

- Reduces the risk of overfitting by bootstrapping and ensemble methods.
- Has good track record for handling imbalanced or high dimensional data.

Disadvantages:

- It does better with classification than with regression.

TABLE 1: TEST METRICS

Logistic Regression	Model	Random Forest
0.78	Training AUC	0.79
0.79	Testing AUC	0.81
0.7206	Precision	0.67
0.176	Recall	0.31
0.29	F1 Score	0.43

TABLE 2:FI

RANDOM FOREST	LOGISTIC REGRESSION
EURIBOR 3M	EURIBOR 3M
EMPLOYMENT RATE	EMPLOYMENT RATE
PREVIOUS OUTCOME - SUCCESS	PREVIOUS CONTACT IN DAYS
MONTH_MAY	MONTH_OCT

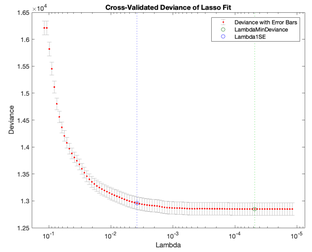

Figure 5

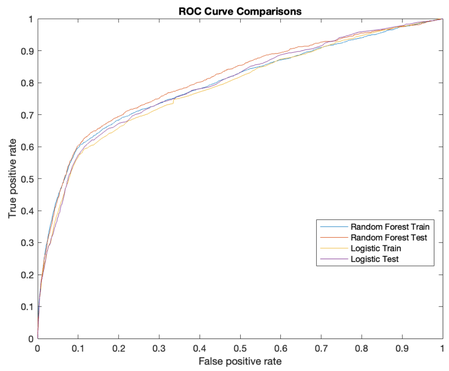

Figure 4

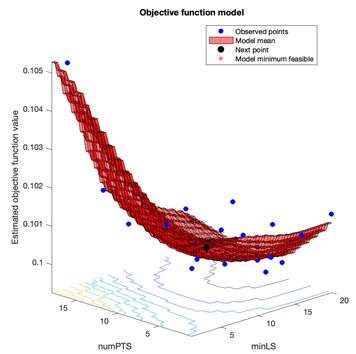

Figure 3: Random Forest Hyperparameter Tuning

- It has been called a 'black box' because of its sheer size due to the large number of deep decision trees produced within it [4].
- Training is computationally intensive.

4. HYPOTHESIS

Previous work on this data set used pre filtering and feature selection to build their models. We wish to compare to results by using models with embedded feature selection. We expect the Random Forest classifier to outperform the logistic regression as it has a higher *average* predictive performance[5]. However Logistic regression has been shown to perform better when there is a balance between the explanatory variables and irrelevant noise variables[6].

5. METHODOLOGY

- Split the data into training and test set through stratified random sampling with 70% training and 30% test.
- For the logistic classifier we apply a 10 fold cross-validation on the training set to approximate the training error. For the random forest we approximate the error through out-of-bag sampling.
- Models are trained iteratively to find the optimal hyper-parameters which minimise training error.
- Our main evaluation metric is the Area Under the Curve (AUC) which provides the advantage of being independent of the class frequency and evaluates the classifier over all threshold values.
- Testing the two best trained models on the test set and comparing their efficiency.

LOGISTIC REGRESSION

PARAMETERS

The model selected is a logit link function with an L1 penalty of 0.0062. We chose the penalty with the lowest average cross validated deviance as shown in Figure 5.

EXPERIMENTAL RESULTS

- We did not find any increase with the training accuracy by adding the L1 penalty. This suggests that although the model was overparameterized, the variation was only due to a few number of predictors.
- Adding the lasso penalty did remove a majority of predictors. Moreover adding a L2—penalty which is not capable of feature selection—removed all our predictors. This suggests that our model had a number of highly irrelevant variables.

RANDOM FOREST

The trained random forest model selected is a bagged ensemble of decision trees with a majority vote as the decision criteria. The model is trained by hyper parameters which were selected through an iterative grid search and Bayesian optimisation.

PARAMETERS

- The number of trees in the forest was fixed at 150
- The number of predictors to sample is 10 at each split
- The minimum, leaf size is 20.

EXPERIMENTAL RESULTS

- Although the literature is indecisive about the optimal number of trees in a random forest, there is some suggestion that there is an absolute upper limit for classification models. The authors[7] suggest a number between 64 and 128 suitable for most classification data sets, after which there are diminishing returns. We chose 150 as a suitable compromise between computational costs and training accuracy.
- After that our cost minimisation problem became more manageable and we used Bayesian optimisation to find the optima function which minimises both: numbers of of predictors to sample and the minimum leaf size as shown in Figure 3. Bayesian optimisation is a sequential method which selects hyper-parameter by prioritising those that appear more promising in the last iteration.
- The number of observations per leaf controls the depth of the forest. Deep trees tend to overfit and shallower ones under-fit. We chose 20 as our upper bound to search.
- The number of predictors to sampled were given a range of 1 through all the predictors.

6. ANALYSIS AND CRITICAL EVALUATION

- The average AUC for both models was lower than that of [2]. However the AUC is not directly comparable as they used a number of different variables than our data set. In particular they used the call duration metric which we decided not to used as it is not available before a marketting call is made.
- However we can compare the relative importance of the variables from their research and ours. Table 2 provides a list of the most important variables for both our models. In particular the daily 3m Euribor rate was shown to be the most significant for both our models and theirs.
- Lasso regularisation in the logistic model did not lead to a better model performance with no significant change in the AUC. Lasso regression biases coefficients to reduce variance and did penalise most of our variable coefficient to zero. The Euribor rate with the largest coefficient by many factors to the next largest one, which is the employment rate. This does imply that most of the variation in our model was because of the economic factors rather than consumer profiling characteristics and does lead give precedence to the idea that the model is ill-advised.
- Ridge recession penalised all our coefficients to zero which does give credence to the idea that we had a large number of highly irrelevant features with only some with moderate effects.
- This is further supported by the relative stability of the test and training AUC of both our models. The low bias and variance of the classification AUC suggest that our models are relatively simple with the prediction variation coming from a limited dimension.
- We posit that this is in large part to the skewed class proportions in out dependent variable. The models are overfitting to the majority class which leads to low accuracy for the minority class. This is quite apparent in Figure 4 where all ROC curves a very similar distribution. Both models show good test AUC scores and good generalisation, but after a certain point the ROC curves become almost linear which suggests the models cannot discriminate well between the two classes above a low threshold.
- This is apparent in Table 1 which provides the test statistics for both our models. The F1 score for the random forest classifier is more than that of the logistic regression model, however most of the discrepancy is because of the higher recall.
- And looking at the ROC curves it is apparent that the random forest classifier has better precision and recall at different thresholds as well. This is probably because the ensemble nature of an RF provides it with a more complication decision rule to pick out the minority class, then a logistic regressor with a linear boundary.
- Or it could be the RF model has overfitted to the noise and it is not apparent because of the low number of samples.

7. LESSONS LEARNT

- Class imbalances lead to inefficient model performance especially with respect to the minority class.
- Logistic Regression requires fine grained tuning in comparison to Random Forest but on the other hand provides better metrics about the model process.
- Embedded feature regularising models are not a catch-all for all problems and it could better to use an iterative method whereby the relevant features are extracted on one model and used to predict on another.

8. FUTURE WORK

- Vary the training size through synthetic sampling methods such as under sampling and SMOTE. This process might provide the model more training data to work with.
- Use alternate cutoffs of predicted probabilities to increase the prediction accuracy of the minority class samples. Alternatively check the suitability of other models which provide tuning of prior probabilities and class weights [3].
- Alternative regularisation techniques such as elastic net.

[1] S. Moro, R. M. S. Laureano, and P. Cortez, 'Using data mining for bank direct marketing: an application of the CRISP-DM methodology', ***undefined*****, 2011, Accessed: Dec. 15, 2021. [Online]**

[2] S. Moro, P. Cortez, and P. Rita, 'A data-driven approach to predict the success of bank telemarketing', ***Decision Support Systems*****, vol. 62, pp. 22–31, Jun. 2014, doi: 10.1016/j.dss.2014.03.001.**

[3] M. Kuhn and K. Johnson, ***Applied Predictive Modeling*****, 1st ed. 2013, Corr. 2nd printing 2018 edition. New York: Springer, 2013.**

[4] H. Zhang and M. Wang, 'Search for the smallest random forest', ***Stat Interface*****, vol. 2, no. 3, p. 381, Jan. 2009.**

[5] R. Couronné, P. Probst, and A.-L. Boulesteix, 'Random forest versus logistic regression: a large-scale benchmark experiment', *BMC Bioinformatics*, vol. 19, no. 1, p. 270, Jul. 2018, doi: 10.1186/s12859-018-2264-5.

[6] K. Kirasich, 'Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets', vol. 1, no. 3, p. 25, 2018.

[7] T. Oshiro, P. Perez, and J. Baranauskas, 'How Many Trees in a Random Forest?', Jul. 2012, vol. 7376. doi: 10.1007/978-3-642-31537-4_13.