Summary Report: A Classification Approach to Predicting Fixed Term Savings Deposits

Ivor Walker

Introduction

Predicting the outcome of a marketing campaign in advance can reduce calls to customers unlikely to subscribe and improve efficiency. The modelling process itself provides insights into factors that influence customer decisions that can be used to improve products and marketing alike. I predict whether a customer makes a deposit in advance of a marketing campaign by fitting three binary classifiers to the bank marketing dataset and evaluate their performance on unseen data.

Methodology

This dataset comprises of categorical, numerical and economic predictors. I bin some categorical and all numerical variables to create a "sensitive" dataset on which a sensitive linear model could be trained. During Exploratory Data Analysis (EDA), I discovered a non-linear time trend 1 and that economic variables varied significantly but non-linearily with the response. I fit an exploratory Generalised Additive Model (GAM) to the sensitive dataset to model these linear and non-linear variables together. Among the linear variables, I found that method of contact, past marketing campaign outcomes, credit default and number of times contacted were significant predictors of success. 1 Among the non-linear variables, I found that the time component and interest rates were significant - the other economic variables were likely not significant as they were determined by interest rates and stage in the business cycle. I applied my chosen models to an "insensitive" dataset, where all numerical variables were presented as continuous and all levels of categorical variables were included, because these models can infer the correct bins themselves. To genuinely evaluate the model's performance on future data, I train the model on the first 260 days (93% of data) and evaluate its performance on predicting remaining 130 days (7% of data). Because the most recent data contains the fewest records 1, this split needs to be far enough forward to give the model some of the most recent temporal patterns but far enough back to create a large enough test set and avoid overfitting. It is a reasonable compromise but still causes some overfitting as the split departs from the 80/20 rule of thumb.

I used Stochastic Gradient Descent (SGD), Decision Tree (DT) and Random Forest (RF) classifiers as they are not computationally intensive to train on large datasets such as ours. They have characteristics that do not change with the dataset (hyperparameters) that can be tuned to improve performance. For each classifier, I defined a "parameter grid" of the possible values of each parameter, trained models on a randomly chosen subset of the training dataset with every combination of hyperparameters, and used the hyperparameters of the best performing model. Tuning models using the entire training set, a larger parameter grid or in an iterative manner (i.e tuning, then changing parameter grid based on values selected in previous tuning) could improve performance but were computationally infeasible. These models produce a probability of success, so after training I choose a threshold to maximise performance on the training set. My measures of performance are the integral approximations of the Receiver Operating Characteristic Curve (integral_ROC) and the F1 score, which I chose because they are robust to the class imbalance in the dataset. F1 measures the performance of a model at its optimal decision threshold, while ROC integral measures the performance of a model across all possible thresholds.

Results

SGD appears to have the best performance but it produces probability predictions between 0.23 and 0.69 yet its decision threshold is 0 - unlike DT and RF, it simply predicts that a customer will subscribe no matter the data, so has failed to learn anything from the training data. RF is less likely to produce false positives than DT, but this is outweighed by its comparative inability to identify when a consumer will subscribe. RF performs best across all thresholds, but DT performs best at the optimal threshold so I select it as the best model so far. SGD fails in the same way within the training data, but RF and DT extremely overfit to the training data. Removing the day_id variable improves the performance of both models on test data, but these models still overfit extremely on training data.

Discussion

SGD selected the least complex parameters possible and selected "modified_huber" as its loss function where I would have expected it to choose "log_loss". Huber is more useful in classification cases where differences between classes are marginal, unlike "log_loss" where differences between classes are more distributed. The high proportion of categorical data may have guided SGD towards a more marginal classification outcome than in a dataset with less categorical data. I declined to tune the Platt scaler wrapped around the SGD due to computational limitations, but creating a parameter grid for it could also widen the probability window. SGD is a simpler model than RF and DT and could benefit from being trained on the simpler "sensitive" dataset but declined to do so due to computational restrictions.

The improvement in performance from removing the day_id indicates that RF and DT are ignoring the temporal component of the data. A temporal component may have been too complex for these models to learn as these models are not designed for time series data.

There is some inherent overfitting in the model due to the split of the training and test data. There may be some confounding variables causing these two models to overfit, but EDA identified no single variable that overinfluenced the process in the sensitive dataset. Training all models on the "sensitive" dataset may reduce overfitting but would require more computational resources than I have available. Identifying the confounding variables would also be useful but require re-tuning and training the models on all possible combinations of columns, so is computationally infeasible.

Conclusion

The best working model is RF with day_id removed, but it is worse than a classifier predicting that all customers will subscribe due to extreme overfitting and an inability to learn the temporal component of the data. A more complex model designed for time series such as a Long Short Term Memory (LSTM) network could learn the temporal component. Training on "sensitive" data or identifying the confounding variables could reduce overfitting. Predictive performance can be improved more generally by improving hyperparameter tuning. However, all these changes would require more computational resources than I have available.

Appendix

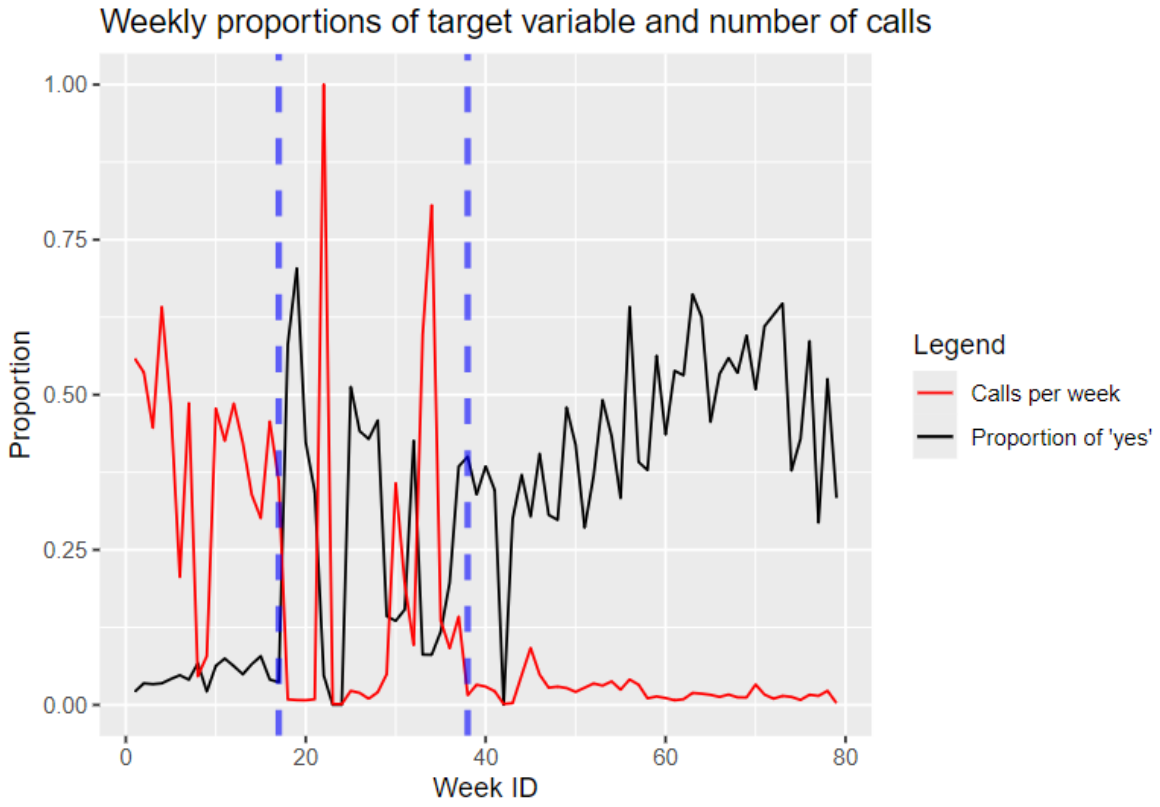

Figure 1: Total calls made per week and the number of successes. There appear to be three segments: weeks 1-17 has a high number of calls and low successes, weeks 18-38 has a medium but unstable number of calls and successes, and after week 39 the number of calls are low but successes are high.

Significant predictor	**Baseline**
Contacted by telephone	Contacted by cellphone
Day is Tuesday	Day is Monday
Day is Wednesday	;
Day is Thursday	;
Day is Friday	;
Customer has defaulted or unknown credit status	Customer has no credit default
Customer has been contacted in a previous campaign	Customer has not been contacted before
Previous campaign failed and customer was contacted once	User was not contacted in a previous campaign
Previous campaign failed and customer was contacted twice	;
Previous campaign failed and customer was contacted thrice	;
Previous campaign failed and customer was contacted four times	;
Customer was contacted over nine times in the campaign	Customer was contacted once
Estimated effect	**Standard error**
0.56	0.04
1.30	0.08
1.44	0.09
1.22	0.08
1.26	0.08
0.84	0.05
4.18	1.19
0.68	0.04
0.46	0.08
0.45	0.16
0.16	0.13
0.72	0.11

Table 1: Summary of statistically significant linear predictors from the exploratory model. The estimated effect is the change in the odds (coverted from log-odds) of success compared to the baseline, e.g customers contacted by telephone are 0.56 times as likely to subscribe compared to customers contacted via cell. The standard error is the uncertainty in this estimate. These predictors have sufficiently large effects and low uncertainty to be considered significant.