IDENTIFYING OUT OF BUSINESS RECORDS ON THE NASS LIST FRAME USING BOOSTED REGRESSION TREES

Gavin R. Corral[1], Andrew J. Dau[2], Jodie M. Sprague[3]

[123]United States Department of Agriculture, National Agricultural Statistics Service, 1400 Independence Avenue SW Washington, DC 20250

Abstract

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) produces hundreds of publications annually. The research conducted at NASS is based on survey data, collected using the NASS list and area frames. Therefore, it is imperative that the NASS list frame is complete and up-to-date in order to produce valid and accurate estimates for agriculture. For this reason, NASS is constantly updating the list frame by adding new farms. Conversely, farms also go out-of-business, and these farms need to be removed from the list frame for it to stay current. In this paper, we examine the efficacy of boosted trees to identify out-of-business records prior to data collection. We found that boosted regression trees outperformed logistic regression and random forests. Boosted regression trees were shown to have the lowest misclassification rate and highest R^2.

Key Words: Boosting, Machine Learning, Regression Trees, Boosted Regression, Supervised Learning

Introduction

The National Agricultural Statistics Service (NASS) releases reports driven by survey data on an annual basis. The quality of those surveys are dependent on up-to-date sampling frames. The sampling frames used by NASS are updated continually with new records being added each year. However, overtime sampling frames age and aging records can create “deadwood”. Deadwood records are shown on the frame as in business, but are actually out of business and need to be updated for the sampling frame to maintain high quality.

Locating the out of business records is a difficult problem and one that is not new to researchers and practitioners. Various methods have been used to estimate the probability of deadwood. One example would be logistic regression. In this case, one would select a few predictor variables and build a model based on R^2 criteria. The issue with this is that many of these sampling frames have dozens or even hundreds of possible predictor variables with sometimes complicated non-linear relationships (Westreich et al. 2010) to the presence or absence of deadwood. Machine learning techniques have been suggested as promising alternatives to logistic regression (Lee et al. 2010, Strobl et al. 2009). McCaffrey et al. (2004) note that machine learning techniques, such as boosted regression trees, are well suited to deal with these obstacles and provide accurate estimates.

The aim of this project is to develop a model to best estimate the probability of a record being deadwood. We used the sampling frame for small grain county estimates from 8 defined regions in the United States. The sampling frame consisted of 50 variables. We used a boots on the ground approach to verify model accuracy. We hope to use it as a tool to update the NASS sampling frame, thereby ensuring quality estimates.

Methods

JMP Pro 12 was used for all analysis in this project. Within JMP, we utilized the Boosted Tree platform. In general, the boosted tree platform produces an additive decision tree model that is a product of many smaller decision trees that are assembled as layers. The tree in each layer consists of a small number of splits, which in our case was a max splits per tree. Each layer is then fit using a recursive fitting methodology (Proust 2016).

The partition platform recursively partitions data according to a relationship among the predictors and the response variable – creating decision trees. The partition algorithm does an exhaustive search of all possible splits of predictors to best predict the response. The partitioning of the data is done recursively to form a tree of decision rules. The splits continue until the desired fit is obtained. Then, the partition algorithm chooses an optimum split from a large number of possible splits.

The objective function (1) is used to optimize the model. The objective function consists of two parts,

$$Obj(\theta) = L(\theta) + \omega(\theta) \qquad (1)$$

the training loss ($L(\theta)$), which measures how well the model fits on the training data and the regularization ($\omega(\theta)$), which measures complexity of the model. Ideally the model fits the training data well and is less complex than other models. Optimizing training loss ensures predictive models. By fitting the training data well, one is hopefully fitting the underlying distribution well, which is assumed to be similar to the training data. Optimizing regularization ensures simple models. A simpler model yields smaller variance in future predictions, resulting in stable distribution of future predictions.

Finally, we applied our model to several different surveys at different years and months. Then field employees contacted or attempted to contact all records listed as deadwood to confirm or refute the model findings.

Results

Eleven variables were chosen for the final model. The resulting R^2 for the model is 0.1529 and the misclassification is 0.0177 (figure 1). The confusion matrix in figure 1 further illustrates the low misclassification as well as the model's tendency to predict deadwood when the record is not actually deadwood.

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.1529	0.1337	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1646	0.1443	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.0750	0.0772	∑ -Log(ρ[j])/n
RMSE	0.1260	0.1279	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0320	0.0325	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0171	0.0177	∑ (ρ[j]≠ρMax)/n
N	392193	97679	n

Confusion Matrix

Training

Actual	Predicted	
deadwood	0	1
0	385172	118
1	6602	301

Validation

Actual	Predicted	
deadwood	0	1
0	95899	46
1	1680	54

Figure 1 Fit statistics for final boosted tree model and confusion matrix illustrating the predictive accuracy of the model

The Receiver Operating Characteristic (ROC) curve measures the sorting efficiency of the models fitted probabilities, which are used sort the response levels. The higher the curve from the diagonal line the better diagnostic ability if the binary classifier. Generally, an area under the curve of 0.70-0.80 is considered “fair”.

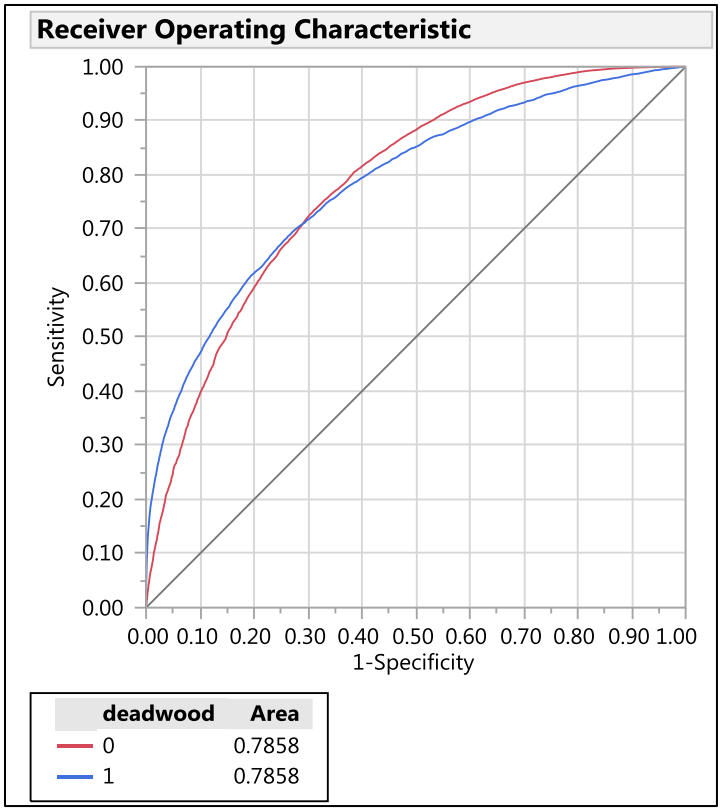

Figure 2 ROC curve. Area under the curve is measure of predictive power of model

Table 1 provides insight into how the model performed. This models was applied to a number of surveys and had field employees attempt to verify each deadwood record. Of those records identified as deadwood, 38% were deadwood and with inaccessible records removed up to 58% of records were shown to be deadwood. As a result, we were able to remove up to 228 deadwood records.

Table 1 Shows a breakdown of records indicated as deadwood by the model.

Survey	Month	Year	Deadwood Removed	Deadwood ID'd	Deadwood %	Inaccessible	Inaccessible %	Deadwood % W/Out Inac
CROPS APS	6	2016	11	35	31.43%	9	25.71%	42.31%
CROPS APS	9	2016	22	76	28.95%	15	19.74%	36.07%
CROPS CE	9	2016	135	356	37.92%	71	19.94%	47.37%
CROPS APS	12	2016	71	294	24.15%	105	35.71%	37.57%
CROPS APS	3	2017	121	355	34.08%	96	27.04%	46.72%
AG LABOR	4	2017	43	128	33.59%	43	33.59%	50.59%
CROPS APS	6	2017	228	600	38.00%	217	36.17%	59.53%
Total			**631**	**1844**	**34.22%**	**556**	**30.15%**	**48.99%**

Conclusion

Using this model, a high rate of deadwood in the list frame was identified. In the first test, four regions is in the NASS list frame were considered. The model was applied to these records with a 20% threshold. That is, a record was flagged as potential deadwood if it had a 20% chance or greater of being deadwood. It was found that 38% of all flagged records were verified by field employees as deadwood and the operation was in fact out of business.

Even with fairly low success rates, this method is more successful than previous methods including having data collectors "estimate" which records were out of business, given their response history. The fit statistics of the model are not ideal, but the model does make for a useful tool at NASS by providing a starting point for cleaning up the sampling frames from deadwood.

Work Cited

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, *29*(3), 337-346.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

Proust, M. (2016). SAS Institute Inc. JMP® Pro 12 Predictive and Specialized Modeling. Cary, NC: SAS Institute Inc.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), 826-833.