Comparison of Ordinal Logistic Regression with Tree-Based Methods in Predicting Socioeconomic Classes in the Philippines

Michael Daniel C. Lucagbo
School of Statistics
University of the Philippines

The task of classifying Philippine households according to their socioeconomic class (SEC) has been tackled anew in a collaborative work between the Marketing and Opinion Research Society of the Philippines (MORES), the former National Statistics Office (NSO) and the University of the Philippines School of Statistics. This new system of classifying Philippine households has been introduced in the 12th National Convention on Statistics, in a paper entitled *1SEC 2012: The New Philippine Socioeconomic Classification*. To predict the SEC of a household, certain household characteristics are used as predictors. The 1SEC Instrument, whose scoring system is based on the ordinal logistic regression model, is then used to predict the household's SEC. Recently, the statistical literature has seen the development of novel tree-based learning algorithms. This paper shows that the ordinal logistic regression model can still classify households better than three popular tree-based statistical learning methods: bootstrap aggregation (or bagging), random forests, and boosting. In addition, this paper identifies which clusters are easier to predict than others.

Keywords: socioeconomic classification, ordinal logistic regression, bagging, random forests, boosting

1. Introduction

The Marketing and Opinion Research Society of the Philippines (MORES), the former National Statistics Office (NSO), and the University of the Philippines School of Statistics have collaborated on a new scheme of classifying Philippine households according to their socioeconomic class (SEC). The *1SEC 2012* study has defined *clusters* or *socioeconomic classes* which were formed as a result of a

sequence of nonhierarchical cluster analyses on households included in the 2009 Family Income and Expenditure Survey (FIES). The 1SEC team came up with nine clusters.

In the 1SEC classification scheme, a household's total family expenditure determines its actual cluster. Because it is next to impossible to collect data on total family expenditure in quick surveys, it is more practical for these surveys that other household information related to expenditure but easier to ask be used in predicting a household's true cluster. Moreover, it is necessary that a good algorithm with superior predictive ability be implemented in classifying households based on these proxy household information.

For this reason, one of the goals of Bersales et al. (2013) is to devise a method of predicting the SEC of a household based on variables which are easy to ask. The variables used to predict SEC should be answerable even to fifteen-year-olds (the standard for "easy to ask"). Thus, only qualitative and count variables are considered. Bersales et al. (2013) trim the list of predictors to 36 variables, all of which are measured in the 2009 FIES. The list can be categorized into nine groups: (1) quality of consumers in the household, (2) number of selected energy-using facilities owned, (3) urban and regional membership, (4) transport type ownership, (5) water source type, (6) connectivity, (7) living space assets, (8) living shell, and (9) tenure of home.

Ordinal logistic regression is run with SEC as response variable and the 36 as explanatory variables. The sizes of the coefficients of the regression run are then used as basis for the scorecard, the MORES 1SEC Questionnaire, which predicts the true SEC of a household. As will be shown later, the ordinal logistic regression model achieves a 41.8% hitrate (percentage of households correctly classified) for exact-cluster predictions of the test set. This hitrate is marvelous, given that there are nine clusters involved in the classification problem. Moreover, the hitrates for four out of the nine clusters are above 50%. The households which the model seemingly finds difficult to classify belong to what may be considered the "middle class."

The performance of the ordinal logistic regression in classifying households is tough to beat. Lucagbo (2015) has attempted to improve on the methodology of classifying households by pitting the ordinal logistic regression model against some state-of-the-art statistical learning methods: artificial neural networks (ANN), support vector machines (SVM), and discriminant analysis. The ANN and SVM methods have resulted in hitrates which are only marginally higher. In terms of neighboring-cluster hitrates (where, for example, a misclassification of a household from the 4^{th} cluster to the 3^{rd} or 5^{th} clusters is still considered correct classification), the ordinal logistic regression model outperforms the others.

The main objective of this paper is to show that ordinal logistic regression still has superior performance in classifying households compared to a different class of statistical learning methods: tree-based methods. In addition, the paper identifies the clusters which are easier to predict than others. The classification methods used in this paper will be referred to as *classifiers*. The percentage of correctly classified households will be called *hitrate* and used as a measure of predictive ability.

2. Review of Related Literature

2.1 The 1SEC classification scheme

The classification scheme of Bersales et al. (2013) was presented in the 12th National Convention in Statistics and the 2012 MORES National Congress held in Plantation Bay, Cebu. The members of the *MORES 1SEC team*, are Lisa Grace S. Bersales, Nicco Emmanuel de Jesus, Luzviminda Barra, Judith Rachel Mercado, Maria Beatrice Gobencion, and this author.

The dataset used by the team is the 2009 FIES. The households of the Autonomous Region of Muslim Mindanao (ARMM) were excluded in the runs. All in all, there were 36,812 households used. These were divided into a training set (a random sample of 90% per region) and a test set (the remaining 10%).

The distribution and median total annual family expenditure by cluster is shown in Table 2.1. The variables which are used for the classification problem of identifying the household's SEC are listed in Table 2.2.

Table 2.1. Median Total Annual Family Expenditure by Cluster

Cluster	Number of Households	Percentage of All Households (%)	Median Total Family Expenditure (in Php)
1	2274	6.18	34,744.50
2	6075	16.50	59,653.00
3	8329	22.63	89,469.00
4	4976	13.52	123,997.50
5	4629	12.57	161,068.00
6	3447	9.36	208,665.00
7	3222	8.75	274,245.00
8	2845	7.73	400,852.00
9	1015	2.76	738,592.00

Table 2.2 List of Variables Needed in Predicting SEC of Households Based on MORES 1SEC Methodology

1.	Type of Place of Residence (Urban or Rural)
2.	Regional group where household is located
3.	Does the household spend on laundry services?
4.	Does the household pay tuition fees in cash?
5.	Does the household spend for maid/boy services?
6.	Does the household on LPG?
7.	Does the household spend on firewood?
8.	Does the household spend on charcoal?
9.	Does the household spend on school service (land and water)?
10.	Does the household spend on air fare transport?
11.	Does the household receive cash receipts, assistance from abroad?
12.	Number of TVs owned by household
13.	Number of airconditioners owned by household
14.	Number of refrigerators owned by household
15.	Number of microcomputers owned by household
16.	Number of washing machines owned by household
17.	Number of stereos owned by household
18.	Number of VCRs owned by household
19.	Number of cars owned by household
20.	Number of motorcycles owned by household
21.	Number of phones owned by household
22.	Number of sala sets owned by household
23.	Highest grade completed by the household head
24.	Household house type of roofing
25.	Household house type of wall
26.	Household house building type
27.	Household house toilet facility
28.	Household head occupation
29.	Employment of spouse of household head
30.	Household tenure status
31.	Household head kind of business
32.	Household main source of water
33.	Marital status of household head
34.	Household type
35.	Number of household members 60 years old and over
36.	Number of employed members in the household

2.2 Review of the classifiers

2.2.1 Cumulative Logit Model

The response variable, socioeconomic class, has nine levels (clusters one to nine). Since the response variable is measured in the ordinal level of measurement, the model to be used should take the ordering of the response categories into account. The study uses the cumulative logit model for ordinal response. Following the discussion of Agresti (2007), which this paper outlines, Y represents the response variable, j=1,2,…9 represents the nine clusters, and π_j the probability that a household belongs to cluster j. The cumulative logits are defined as

$$\text{logit}\left[P\left(Y \le j\right)\right] = \log\left[\frac{P\left(Y \le j\right)}{1 - P\left(Y \le j\right)}\right] = \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right], \quad j = 1, ..., J-1$$

As Agresti (2007) describes it, a model for cumulative logit j is similar to a binary logistic regression model in which the categories from 1 to j combine to form a single category, while categories (j + 1) to J form a second category. For an explanatory variable, the model

$$\text{logit}\left[P\left(Y \le j\right)\right] = \alpha_j + \beta x, j = 1, ..., J-1$$

has parameter β describing the effect of x on the logarithm of the odds of response in category j or below. Since the parameter does not have a j subscript, the model is identical for all the (J–1) cumulative logits, enabling us to have a single parameter, instead of (J–1), that describes the effect of x. It can be shown that for every unit increase in x, the odds of response below any given category multiplies by e^{β}.

2.2.2 Bagging

James et al. (2013) explain that bootstrap aggregation, or bagging, is a general-purpose procedure where the bootstrap is used in a new context: to reduce the variance of a statistical learning method. The idea is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions since taking many training sets, performing predictions for each, and then averaging the predictions, is a natural way of reducing the variance.

Since generally one does not have multiple training sets, one option is to bootstrap by taking repeated samples from the same training set. Thus, B different bootstrapped training sets are generated. In the b^{th} bootstrapped training set, the method is trained to arrive at the predicted response $\hat{f}^{*b}(x)$, and finally average the predictions across the bootstrapped training sets to obtained the prediction based on bagging:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Bagging can be applied to predict a qualitative outcome variable Y (as is needed in this study). The simplest approach is by taking the majority vote: the overall prediction that occurs most commonly among the B predictions.

2.2.3 Random forests

Random forests can provide improvements over bagging by *decorrelating* the trees. James et al. (2013) explain that when building the decision trees, each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those *m* predictors. A different sample of *m* predictors is taken at each split, where typically $m \approx \sqrt{p}$.

The rationale behind this is that if there are strong predictors, most of the trees will use the strong predictor in the top split, thus making the bagged trees correlated with each other. Averaging highly correlated quantities does not lead to as large of a reduction in variance as averaging uncorrelated quantities. By considering only a subset of predictors at each split, $(p - m)/p$ of the splits will not consider the strong predictors. This *decorrelates* the trees, reducing the variance of the average prediction.

2.2.4 Boosting

James et al. (2013) describe that boosting works in a way similar to bagging, except that the trees are grown sequentially: each tree is grown using information from previously grown trees. Boosting does not involve bootstrap sampling; instead, each tree is fit on a modified version of the original dataset. The boosting approach is said to *learn slowly*, instead of *fitting the data hard* by fitting a single large decision tree.

2.3 Applications of the tree-based algorithms

In what follows, the merits of the tree-based learning algorithms examined in this paper are illustrated in various applications.

Breiman (1996a) applies the bagging method to classification trees using several datasets from diverse fields. The results show huge decreases in the average misclassification rates (ranging from 20% to 47%) when bagging is used, instead of using just a single decision tree. Breiman (2001) cites bagging as an example in saying that "significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class."

Debeir et al. (2002) use bagging for digital land-cover and land-use classification. They compare the performances of three classifiers: K-Nearest Neighbors, the C4.5 decision tree classifier introduced by Quinlan (1993), and the BAGFS classifier of Latinne (2003), which combines bagging with multiple feature subsets. The results show that BAGFS gives the best overall accuracy. Machova et al. (2006) apply bagging to data from the Reuters 21578 collection of documents, as well as documents from an Internet portal of TV broadcasting company Markiza. They affirm the suitability of bagging for increasing the efficiency of standard machine learning algorithms.

Kudo et al. (2004) use the boosting algorithm with the decision stump classifier as the "weak learner" of boosting in classifying cellphone reviews (positive or negative) and also classifying chemical compounds by carcinogenicity. In predicting the chemical compound, the boosting algorithm outperforms Support Vector Machines (SVM) at a statistically significant level, leading the authors to conclude that "the boosting algorithm is accurate and efficient for classification tasks involving discrete structural features."

Rottman et al. (2005) use boosting to classify different indoor environments (e.g., kitchens, offices, seminar rooms) into semantic classes. The algorithm is then implemented and tested using a mobile robot equipped with a laser range finder and a camera system. The experiments carried out show that the algorithm performs well in classifying indoor environments.

The AdaBoost is a boosting algorithm which Bauer and Kohavi (1998) talk about when they give what they believe is "a more realistic view of the performance improvement one can expect." Specifically, they say that AdaBoost gives an average error rate reduction of 24% over the Naïve Bayes Classifier.

Schapire and Singer (1999) describe several improvements of the AdaBoost and also some generalizations to multiclass problems. They conclude that the experimental results with the improved boosting algorithms show that dramatic improvements in training error are possible when a fairly large amount of data is available. Breiman (1996b) tested some well-known data sets and showed that classification and regression trees (CART) plus AdaBoost did significantly better than any of the commonly used classification methods.

A couple of studies which use both bagging and boosting are now cited. Pal (2007) evaluates the performance of bagging and boosting for remote sensing classification, and concludes that when there is no noise in the training data, bagging and boosting both show an increase of about 3 to 4% of classification accuracy in comparison to the accuracy achieved with a univariate decision tree classifier. When there is noise in the data, Pal (2007) observes that bagging works well but boosting is severely affected.

Bandi and Srihari (2005) illustrate an interesting application of bagging and boosting. They classify handwriting into a writer's demographic category (in particular, gender, age, and handedness of a writer). Bagging and boosting are applied in neural networks. The results show that the accuracies achieved by boosting are significantly higher than what has been observed before for demographic classification.

Breiman (2004) describes random forests as "an accurate algorithm having the unusual ability to handle thousands of variables without deletion or deterioration of accuracy." In an earlier paper, Breiman (2001) argues that the use of the strong law of large numbers assures us that overfitting is not a problem for random forests. In addition, Breiman (2001) argues that the accuracy of a random forest depends on the strength of the individual trees and a measure of the dependence between them.

Chehata et al. (2009) use random forests for urban classification (whether buildings, vegetation, natural ground, or artificial ground). They claim that random forests provide an accurate classification and run efficiently on large datasets. In their study, random forests classification using selected variables provides an overall accuracy of 94.35%. Breiman (2001) experiments with a simulated dataset with a training sample size of 1000 and a test set of 4000. The result shows that random forests compare with the Bayes error rate in terms of accuracy. Bosch et al. (2007) use random forests for image classification (classifying an image by the object category that it contains). They demonstrate that using random forests with an appropriate node test reduces training and test costs significantly over support vector machines, and shows comparable performance.

3. Methodology

The entire 2009 FIES data set was divided into a training set and a test set. The training set was obtained by taking 90% of the households per region. There are 36,812 households in the entire 2009 FIES dataset (excluding households from the ARMM). Of the households, 33,138 were assigned to the training set and 3,674 to the test set.

The results of the analysis are shown in terms of confusion matrices for each of the classifiers. The *confusion matrix* is a cross-classification between a household's actual cluster and its predicted cluster. The confusion matrix for both the training and test sets will be shown, and a comparison of the hitrates of the classifiers will be made afterwards.

The variables included as predictors of socioeconomic classification are either categorical variables (measured in the nominal or ordinal levels) or count variables. These variables are listed in Table 2.2. Variables 1 and 33 are binary variables. Variables 2, 23 to 32, and 34 are categorical variables with more than two categories. Variables 3 to 11 ask information on household expenses. These variables were originally continuous in the 2009 FIES questionnaire but have been transformed to binary variables in the 1SEC Instrument, and are therefore treated as dummy variables in this study. Variables 12 to 22, 35, and 36 ask about household facilities and characteristics of household members; they are count variables. The categorical variables with more than two categories have been transformed to several dummy variables for inclusion in the runs. There are a total of 61 predictors once these transformations of the categorical variables are included.

For the ordinal logistic regression model, a household's predicted probability of inclusion was computed for each of the nine clusters. Each household was then classified as belonging to the cluster with the highest predicted probability. The R package 'rms' was used in running the ordinal logistic regression model.

For the bagging algorithm, the *node size*, or minimum number of observations for each terminal node was set at 30. Larger settings of the node size make for

smaller trees. All of the 61 predictors were considered for each split of a tree in bagging. On the other hand, for random forests, the node size was still set at 30. However, to decorrelate the trees, the number of split candidates was reduced to $m = \sqrt{61} \approx 8$. The R package 'randomForest' was used to implement both bagging and random forests.

For the boosting algorithm, a total of 220 trees for fitting was found sufficient. Since there are nine clusters, the assumed distribution for the response variable is the multinomial distribution. The depth of each tree was limited. The maximum depth of variable interactions was set at 4. The R package 'gbm' was used to implement the boosting algorithm.

4. Results

The confusion matrices of the four classifiers for both the training and test sets are given in Tables 4.1.1 to 4.4.2. The classifiers have relatively high hitrates for households in the 2nd, 3rd, 8th, and 9th clusters. There are differences among classifiers, however, in the hitrates for these clusters. For example, the random forest algorithm can classify households in the 3rd cluster better than the other algorithms (with an impressive 66.3% test-set hitrate), while boosting can predict households in the 9th cluster better than the others (63.4% test-set hitrate).

The classifiers fail to achieve hitrates above 40% for the 4th, 5th, 6th and 7th clusters, which taken together may be considered the "middle class." Moreover, it is the households in the 4th cluster which the classifiers consistently find hardest to classify correctly. Interestingly, for the boosting algorithm, only a paltry 0.2% of training-set and 0.6% of test-set households in the 4th cluster are correctly classified. These classification rates are even worse than chance (1 in 9). A summary of the hitrates per cluster is given in Table 4.5.

Table 4.1.1. Cross Classification based on Ordinal Logistic Regression of Actual and Predicted 1SEC Cluster for the Training Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	518	1359	169	0	1	0	0	0	0	2047	25.3%
	2	165	3070	2087	110	35	0	1	0	0	5468	56.1%
	3	26	1604	4533	768	508	46	10	2	0	7497	60.5%
	4	3	252	2023	819	1087	225	62	8	0	4479	18.3%
	5	0	50	947	697	1531	581	317	44	0	4167	36.7%
	6	0	11	246	259	1011	684	695	193	4	3103	22.0%
	7	0	2	68	109	511	550	1060	588	13	2901	36.5%
	8	0	0	10	19	122	206	661	1341	202	2561	52.4%
	9	0	0	2	2	4	13	51	347	496	915	54.2%
	Total	712	6348	10085	2783	4810	2305	2857	2523	715	33138	42.4%

Table 4.1.2. Cross Classification based on Ordinal Logistic Regression of Actual and Predicted 1SEC Cluster for the Test Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	56	146	25	0	0	0	0	0	0	227	24.7%
	2	20	342	235	9	1	0	0	0	0	607	56.3%
	3	4	186	496	91	49	5	1	0	0	832	59.6%
	4	0	38	204	81	134	29	11	0	0	497	16.3%
	5	0	8	96	92	170	48	45	3	0	462	36.8%
	6	0	1	34	24	116	81	74	14	0	344	23.5%
	7	0	1	6	12	70	52	112	68	0	321	34.9%
	8	0	0	5	2	16	26	66	145	24	284	51.1%
	9	0	0	0	0	0	3	6	37	54	100	54.0%
	Total	80	722	1101	311	556	244	315	267	78	3674	41.8%

Table 4.2.1. Cross Classification based on Bagging of Actual and Predicted 1SEC Cluster for the Training Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	900	952	189	3	3	0	0	0	0	2047	44.0%
	2	372	2807	2135	87	61	4	1	1	0	5468	51.3%
	3	90	1609	4603	524	544	92	29	6	0	7497	61.4%
	4	9	294	2241	575	962	263	112	23	0	4479	12.8%
	5	3	81	1216	537	1336	513	386	94	1	4167	32.1%
	6	1	27	371	248	947	577	667	261	4	3103	18.6%
	7	0	3	134	107	541	497	914	690	15	2901	31.5%
	8	0	2	28	20	164	197	597	1412	141	2561	55.1%
	9	0	0	0	4	9	6	44	410	442	915	48.3%
	Total	1375	5775	10917	2105	4567	2149	2750	2897	603	33138	40.9%

Table 4.2.2. Cross Classification based on Bagging of Actual and Predicted 1SEC Cluster for the Test Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	86	106	34	0	1	0	0	0	0	227	37.9%
	2	46	301	249	8	3	0	0	0	0	607	49.6%
	3	9	195	511	58	49	6	4	0	0	832	61.4%
	4	1	36	234	65	101	37	22	1	0	497	13.1%
	5	0	10	142	60	133	64	45	8	0	462	28.8%
	6	0	6	39	31	101	60	78	29	0	344	17.4%
	7	0	0	10	9	70	60	91	79	2	321	28.3%
	8	0	0	7	4	18	32	68	139	16	284	48.9%
	9	0	0	0	0	2	1	3	40	54	100	54.0%
	Total	142	654	1226	235	478	260	311	296	72	3674	39.2%

Table 4.3.1. Cross Classification based on Random Forests of Actual and Predicted 1SEC Cluster for the Training Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	804	1040	197	3	3	0	0	0	0	2047	39.3%
	2	278	2885	2217	25	58	3	1	1	0	5468	52.8%
	3	55	1633	4926	218	579	45	33	8	0	7497	65.7%
	4	7	275	2552	263	1096	143	121	22	0	4479	5.9%
	5	0	70	1451	280	1574	312	374	106	0	4167	37.8%
	6	1	17	495	113	1137	369	658	313	0	3103	11.9%
	7	0	4	171	48	677	325	877	790	9	2901	30.2%
	8	0	1	27	13	187	138	537	1557	101	2561	60.8%
	9	0	0	3	1	7	8	33	479	384	915	42.0%
	Total	1145	5925	12039	964	5318	1343	2634	3276	494	33138	41.2%

Table 4.3.2. Cross Classification based on Random Forests of Actual and Predicted 1SEC Cluster for the Test Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	79	119	29	0	0	0	0	0	0	227	34.8%
	2	31	321	248	3	4	0	0	0	0	607	52.9%
	3	4	183	552	24	61	5	3	0	0	832	66.3%
	4	1	38	269	27	129	13	18	2	0	497	5.4%
	5	0	9	170	34	164	31	41	13	0	462	35.5%
	6	0	3	54	11	117	52	76	31	0	344	15.1%
	7	0	0	16	3	81	46	88	86	1	321	27.4%
	8	0	0	6	2	23	21	62	162	8	284	57.0%
	9	0	0	0	0	0	0	7	48	45	100	45.0%
	Total	115	673	1344	104	579	168	295	342	54	3674	40.6%

Table 4.4.1. Cross Classification based on Boosting of Actual and Predicted 1SEC Cluster for the Training Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	783	988	272	0	3	1	0	0	0	2047	38.3%
	2	351	2853	2180	0	41	34	4	5	0	5468	52.2%
	3	119	2091	4606	10	292	292	31	56	0	7497	61.4%
	4	25	672	2561	11	433	534	128	110	5	4479	0.2%
	5	12	355	1781	11	649	742	284	329	4	4167	15.6%
	6	7	153	731	5	502	690	422	584	9	3103	22.2%
	7	3	76	358	3	396	474	519	1048	24	2901	17.9%
	8	0	15	102	1	150	193	318	1670	112	2561	65.2%
	9	0	4	6	0	12	8	32	513	340	915	37.2%
	Total	1300	7207	12597	41	2478	2968	1738	4315	494	33138	36.6%

Table 4.4.2. Cross Classification based on Boosting of Actual and Predicted 1SEC Cluster for the Test Set

		Predicted Cluster									Total	Hitrate
		1	2	3	4	5	6	7	8	9		
Actual Cluster	1	83	109	35	0	0	0	0	0	0	227	36.6%
	2	42	311	247	0	3	3	0	1	0	607	51.2%
	3	14	239	520	1	24	29	3	2	0	832	62.5%
	4	5	85	261	3	52	60	14	17	0	497	0.6%
	5	2	48	198	0	64	87	32	31	0	462	13.9%
	6	1	18	84	1	63	60	56	61	0	344	17.4%
	7	0	9	33	0	47	63	52	116	1	321	16.2%
	8	0	1	15	0	25	15	35	180	13	284	63.4%
	9	0	1	2	0	3	0	4	49	41	100	41.0%
	Total	147	821	1395	5	281	317	196	457	55	3674	35.8%

Table 4.5. Training and Test Set Hitrates by Classifier and by Cluster

	Cluster	Logistic Regression	Bagging	Random Forests	Boosting
Training Set	1	25.3%	44.0%	39.3%	38.3%
	2	56.1%	51.3%	52.8%	52.2%
	3	60.5%	61.4%	65.7%	61.4%
	4	18.3%	12.8%	5.9%	0.2%
	5	36.7%	32.1%	37.8%	15.6%
	6	22.0%	18.6%	11.9%	22.2%
	7	36.5%	31.5%	30.2%	17.9%
	8	52.4%	55.1%	60.8%	65.2%
	9	54.2%	48.3%	42.0%	37.2%
Test Set	1	24.7%	37.9%	34.8%	36.6%
	2	56.3%	49.6%	52.9%	51.2%
	3	59.6%	61.4%	66.3%	62.5%
	4	16.3%	13.1%	5.4%	0.6%
	5	36.8%	28.8%	35.5%	13.9%
	6	23.5%	17.4%	15.1%	17.4%
	7	34.9%	28.3%	27.4%	16.2%
	8	51.1%	48.9%	57.0%	63.4%
	9	54.0%	54.0%	45.0%	41.0%

Table 4.6 shows the hitrates for the training and test sets for all the four classifiers. As expected, the predictive abilities of the classifiers are consistently better for the training sets than for the test sets. Overall, ordinal logistic regression performs better than the tree-based methods, though its performance is comparable with that of random forests and bagging. Moreover, boosting appears to give the inferior prediction rates.

Table 4.6. Hitrates for Training and Test Sets by Classifier

Classifier	Training	Test
Ordinal Logistic Regression	42.4%	41.8%
Bagging	40.9%	39.2%
Random Forests	41.2%	40.6%
Boosting	36.6%	35.8%

5. Conclusion

The results show that the performance of ordinal logistic regression is superior to the tree-based methods for both the training and test sets. The study thus justifies the use of the ordinal logistic regression model in classifying households. Nonetheless, the performances of bagging and random forests are only slightly inferior to ordinal logistic regression, and offer a useful alternative. Lastly, boosting seems to show poorest predictive ability among the classifiers considered.

The summary of training and test-set hitrates also shows that some clusters are easier to predict than others. In particular, the classifiers have high hitrates for households in the 2nd, 3rd, 8th, and 9th clusters, and all except ordinal logistic regression have relatively high hitrates for the 1st cluster as well. This result seemingly suggests that, regardless of the classifier used, households in the middle class are harder to classify than rich or poor households.

Some recommendations for future studies are in order here. First, although the ordinal logistic regression model has already been compared with ANN, SVM, and discriminant analysis by Lucagbo (2015), future studies can still compare it with other classes of learning algorithms (e.g., Bayesian algorithms and instance-based algorithms). Second, it is worth identifying the subset of predictors which account for much of the predictive abilities of the classifiers. Lastly, the predictive ability of the ordinal logistic regression model given the 1SEC predictors should be continuously investigated for future FIES years.

Acknowledgements

The author would like to thank Dr. Lisa Grace S. Bersales and Mr. Nicco Emmanuel de Jesus for involving him in the MORES 1SEC study. The author also thanks Dr. Joselito C. Magadia for sharing his knowledge of statistical learning methods. Lastly, much gratitude is due to the anonymous referee for opening the author's eyes to the defects of this paper's original manuscript.

REFERENCES

AGRESTI, A., 2007, *An Introduction to Categorical Data Analysis*, 2nd Ed., Hoboken, New Jersey: John Wiley & Sons, Inc.

BANDI, K. and SRIHARI, S.N., 2005, Writer Demographic Classification using Bagging and Boosting, Proc. 12th Int. Graphonomics Society Conference, pp. 133–137.

BAUER, E. and KOHAVI, R., 1998, An empirical comparison of voting classification: bagging, boosting, and variants, *Machine Learning*, 36:105-139.

BERSALES, L.G.S, DE JESUS, N., BARRA, L., MERCADO, J.R., GOBENCION, M.B., and LUCAGBO, M.D., 2013, 1SEC 2012: The New Philippine Socioeconomic Classification, 12th National Convention on Statistics.

BOSCH, A., ZISSERMAN, A., and MUNOZ, X., 2007, Image Classification Using Random Forests and Ferns, 11th International Conference on Computer Vision, pp. 1-8.

BREIMAN, L., 1996a, Bagging Predictors, *Machine Learning*, 24(2):123-140.

_____, 1996b, Bias, Variance, and Arcing Classifiers, Technical Report No. 460, Statistics Department, University of California Berkeley. Available at: www.stat.berkeley.edu.

_____, 2001, Random forests, *Machine Learning*, 45(1):5-32.

_____, 2004, Consistency for a Simple Model of Random Forests, Technical Report No. 670, Department of Statistics, University of California, Berkeley. Available at: https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf.

CHEHATA, N., GUO, L., and MALLET, C., 2009, Airborne Lidar Feature Selection for Urban Classification Using Random Forests, in Bretar, F., Pierrot-Deseilligny, M., Vosselman, G., eds., Laser Scanning 2009, IAPRS, Vol. XXXVIII, Part 3/W8. Available at: http://www.isprs.org/proceedings/XXXVIII/3-W8/papers/p207.pdf.

DEBEIR, O., VAN DEN STEEN, I., LATINNE, P., VAN HAM, P., and WOLFF, E., 2002, Textual and Contextual Land-Cover Classification Using Single and Multiple Classifier Systems, *Photogrammetric Engineering & Remote Sensing*, 28(6):597-605.

JAMES, G., WITTEN, D., HASTIE, T., and TIBSHIRANI, R., 2013, Introduction to Statistical Learning with Applications in R, New York: Springer.

KUDO, T., MAEDA, E., and MATSUMOTO, Y., 2004, An Application of Boosting to Graph Classification, Proceedings of NIPS 2004.

LATINNE, P., DEBEIR, O., and DECAESTECKER Ch., 2000, Different ways of weakening decision trees and their impact on classification accuracy, Proceedings of the First International Workshop of Multiple Classifier System (MCS'2000).

LUCAGBO, M., 2015, Predicting Socioeconomic Classification in the Philippines: Beyond the Ordinal Logistic Regression Model, *The Philippine Statistician* 64(1):1-14.

MACHOVA, K., BARCAK, F., and BEDNAR, P., 2006, A Bagging Method using Decision Trees in the Role of Base Classifiers, *Acta Polytechnica Hungarica*, 3(2):121-132.

PAL, M., 2007, Ensemble learning with decision tree for remote sensing classification, *World Academy of Science, Engineering and Technology*, 2007(1):12-23.

QUINLAN, J.R., 1993, *C4.5: Programs for Machine Learning,* California: Morgan Kaufman Publishers.

ROTTMAN, A., MOZOS, O.M., STACHNISS, C., and BURGARD, W., 2005, Semantic place classification of indoor environments with mobile robots using boosting, Proceedings of the National Conference on Artificial Intelligence.

SCHAPIRE, R.E. and SINGER, Y., 1999, Improved boosting algorithms using confidence-related predictions, *Machine Learning* 37(3):297-336.