ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Prediction of Rich People through Classification Methods

Hyeuk Kim[1]

Assistant Professor, Division of Global Management Engineering, Hoseo University, Asan, Korea[1]

ABSTRACT: Many techniques have been designed as the artificial intelligence is being developed. Newer method has a tendency to beat older method in performance. However, many factors including selection of the model affect the performance of classification. We pursue the best prediction for open census data in the research. We modify variables and choose the best model among many tested models to achieve our research purpose. In the paper, we suggest a guideline for the correct classification including transformation of the variables and selection of the classification method.

KEYWORDS: Classification, Decision tree, Regression, Neural networks, Ensemble.

I. INTRODUCTION

Prediction through classification is the widely-used method in real world [1]. People always make a decision or a prediction whenever they encounter an event. Regression model, linear discriminant analysis [2], and so on are used in the traditional statistics. However, new techniques have been developed as the computer is upgraded and the machine learning area is developed. Decision tree [3], neural networks [4], support vector machine [5] are the single classification methods in machine learning, and the combination of multiple classifications are developed afterwards. Bagging [6], boosting [7], random forests [8] are applications of the combination of multiple classifications which are called ensemble. It is very important for a researcher to make a model carefully [9]. The performance becomes worse with recent advanced classification method for classification unless he treats the variables in data properly. We present the procedure to handle variables and compare several classification methods with real data set. It is the extracted data from US census. The paper consists of four sections. The first section is to introduce the topic and the related areas. The second section describes data which are used for research. we make experiments with multiple classification methods and compare the performances in the third section. In the final section, we make a conclusion and suggest interesting topics for future study.

II. DATA DESCRIPTION

We choose the open data set for reproducibility. The data set is available on UCI Machine Learning Repository [10]. This is very famous website in machine learning community. It supports almost 400 data sets and are maintained by the department of information and computer sciences in University of California, Irvine. Adult data set which we handle in the research is the census income data set. Barry Becker has extracted it from US census database in 1994 [11]. It consists of one target variable and fourteen exploratory variables which are described in Table 1.

Table 1. Description of Adult data set

Variable name	Description	Variable name	Description
workclass	Type of job	education	Level of education
martial_status	Status of marriage	occupation	Occupation
relationship	Family relationship	race	Race
sex	Gender	native_country	Native country
age	Age	fnlwgt	Final weight

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

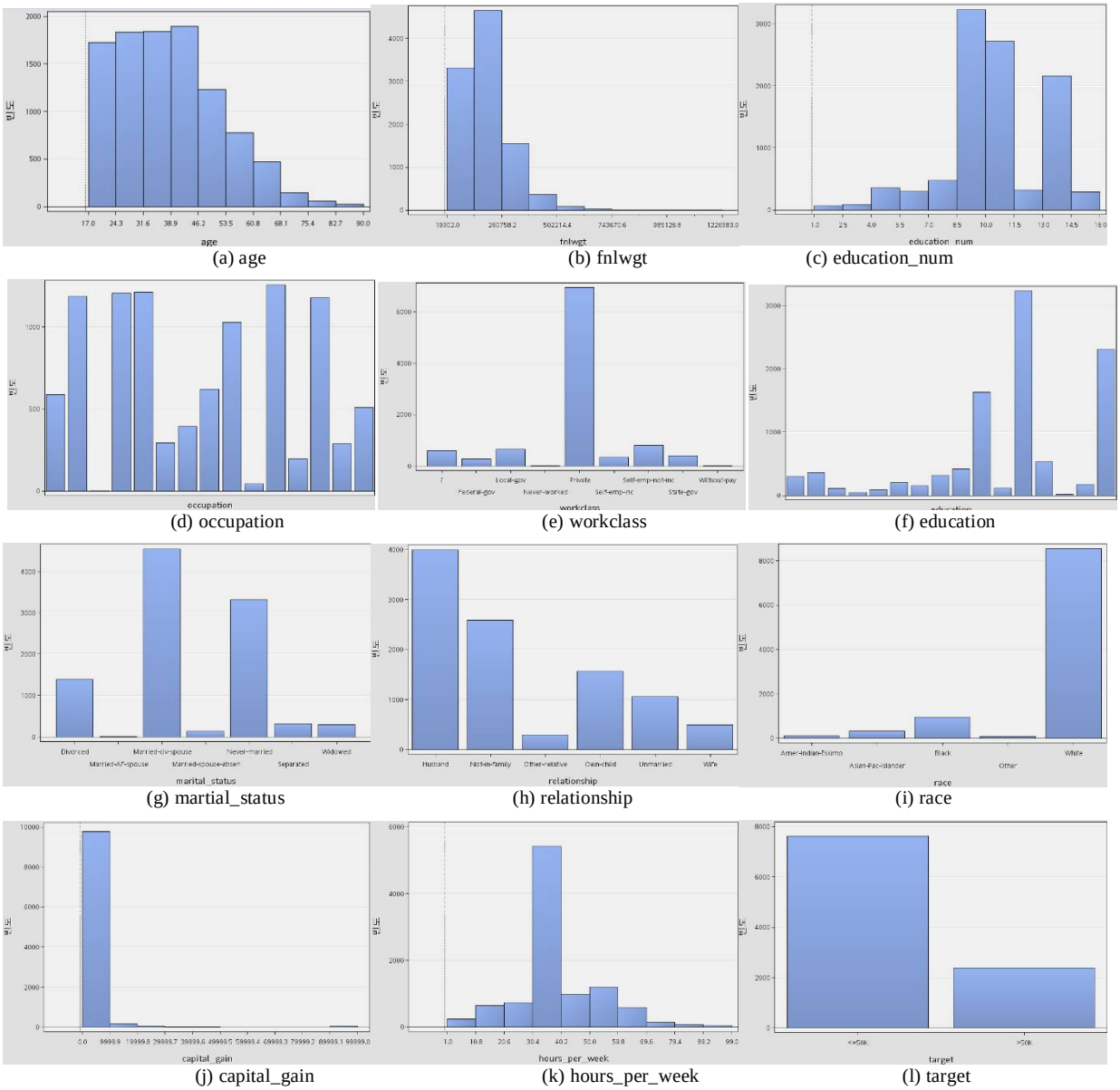
(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

education_num	Total years of education	capital_gain	Capital profit
capital_loss	Capital loss	hours_per_week	Working hours per week
target	Target variable		

We need extra explanation about 'fnlwgt' variable. It is the weight calculated by Population Division at the US census bureau. They use three sets for calculations and mix the values from three sets. the sets consists of a single cell estimate of the population over 16 for each state, controls for Hispanic origin by age and sex, and controls by race, age, and sex. Therefore, people with similar demographic characteristics have similar weights. The target variable has two possible values. '>50K' is the status that the salary is over $50,000 and '<=50K' is the status that the salary is equal to or less than $50,000. The type of some exploratory variables is categorical and the type of others is continuous. The distributions of the variables are described in Figure 1.



(a) age (b) fnlwgt (c) education_num

(d) occupation (e) workclass (f) education

(g) martial_status (h) relationship (i) race

(j) capital_gain (k) hours_per_week (l) target

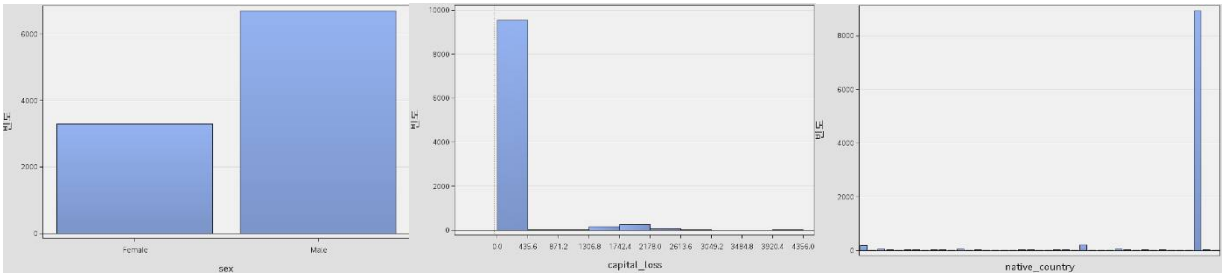
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

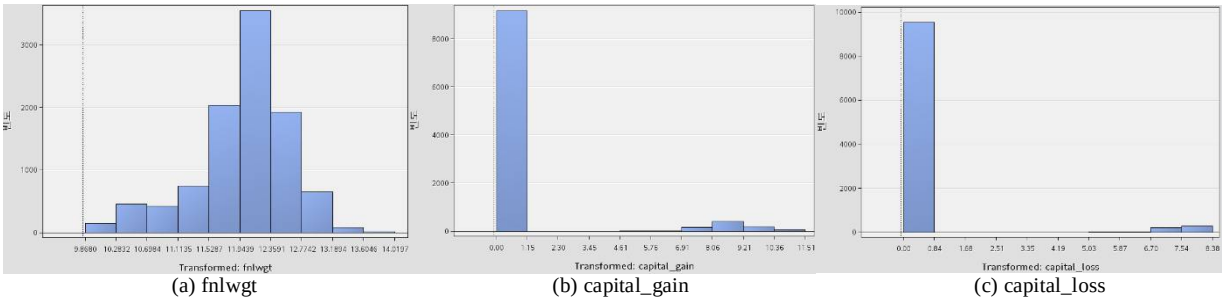
Vol. 6, Issue 8, August 2017



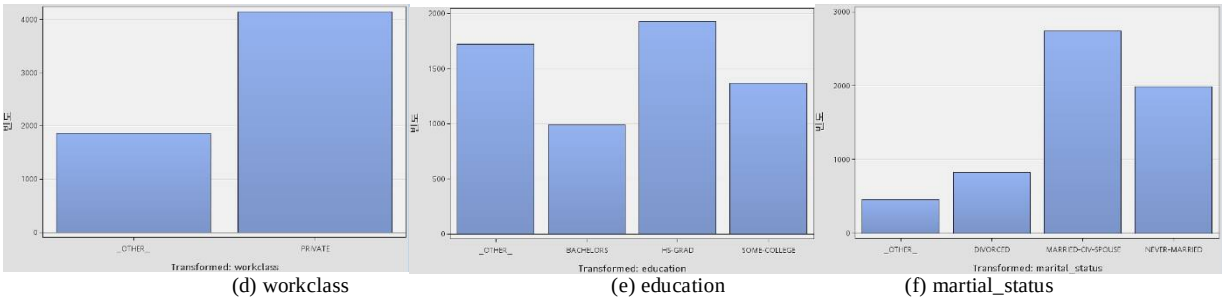
(m) sex (n) capital_loss (o) native_country

Fig. 1. Distributions of the variables before transformation

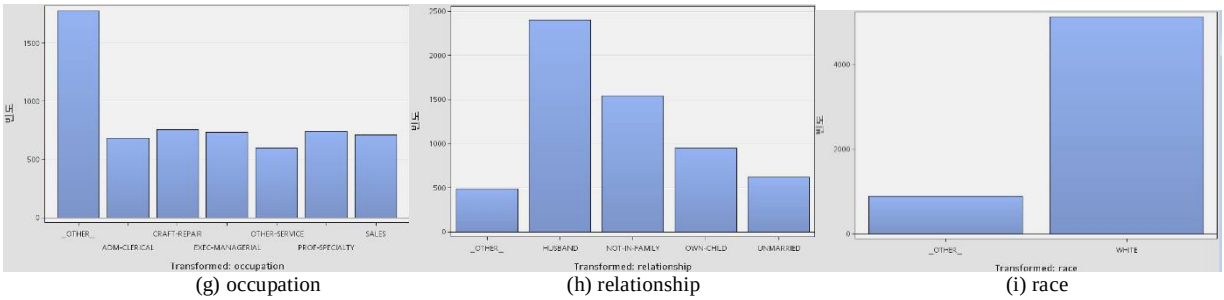
The distributions of many variables are skewed. We need to transform the variables for better prediction. The object of the transformation is to make the values of the variable normalized or simplified. Therefore, we use the logarithm transformation and the group rare levels transformation which are applied to the appropriate variables. The logarithm transformation is applied to fnlwgt, capital_gain, and capital_loss variables. The group rare levels transformation is applied to workclass, education, martial_status, occupation, relationship, race, and native_country variables. The remaining variables are used without transformation. Figure 2 describes the distributions of the transformed variables.



(a) fnlwgt (b) capital_gain (c) capital_loss



(d) workclass (e) education (f) martial_status



(g) occupation (h) relationship (i) race

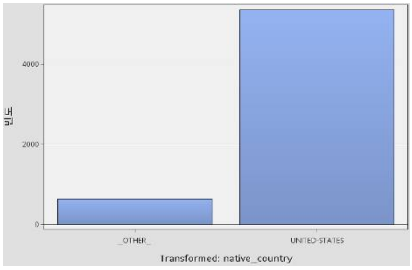
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017



(j) native_country

Fig. 2. Distributions of the variables after transformation

We find that there are three types of change after transformation from Figure 2. The first type of change is that the distribution of the variable becomes normalized. fnlwgt, martial_status, and relationship variables belongs to the first type. The second type of change is that the possible values of the variable are simplified. workclass, education, occupation, race, and native_country variables are in the second types of change. The last type of change is that the effect of change is trivial. The values of capital_gain and capital_loss variables are not changed much even though the log transformation is applied.

III. EXPERIMENTAL RESULTS

We apply several classification methods for the modified Adult data set. Three base classification methods are decision tree, regression, and neural networks. Two derived methods are applied from each base classification method. They are the ensemble approaches such as bagging and boosting. Finally, we consider the combinations of the two base classification methods. The combined methods are the combination of regression and neural networks, the combination of neural networks and decision tree, the combination of regression and decision tree, and the combination of regression, neural networks, and decision tree. Table 2 describes the results of decision tree and the derived classification methods. Bold number means the best performance for the corresponding measure.

Table 2. classification results by decision tree

Classification method	Top 10% cumulative response rate	AUC	KS statistic	Error rate
Decision tree	**98.875**	0.889	0.587	**0.141**
Decisiontree(boosting)	99.387	**0.912**	**0.658**	0.448
Decisiontree(bagging)	98.011	0.895	0.624	0.152

We explain ROC curve [12] before describing AUC. ROC is the abbreviation of a receiver operating characteristic. ROC curve is the curve on the plot of which x-axis is a false positive rate and y-axis is a true positive rate. there is the area from x-axis and to curve which is called AUC. AUC is the abbreviation of an area under curve. There is a rough description about the performance of AUC.

Table 3. performance explanation of AUC

AUC value	Description
0.90-1.00	Excellent
0.80-0.90	Good
0.70-0.80	Fair
0.60-0.70	Poor
0.50-0.60	Fail

KS statistic means Kolmogorov-Smirnov statistic [13] which measures performance of classification models. It measures how the positive and negative distributions are separated. We consider several performance measures simultaneously to decide the performance of each method since there is no best unique measure. Decision tree and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

decision tree with boosting show the better performance than decision tree with bagging. We select decision tree as the best model because the error rate in decision tree with boosting is too bad.

Table 4. classification results by regression

Classification method	Top 10% cumulative response rate	AUC	KS statistic	Error rate
Regression	**90.593**	**0.899**	0.636	**0.157**
Regression(boosting)	88.957	0.892	0.623	0.285
Regression(bagging)	89.980	0.898	**0.637**	0.159

Table 4 describes the results from the regression-type classification which is a traditional approach in Statistics. We easily select a single regression method as the best model.

Table 5. classification results by neural networks

Classification method	Top 10% cumulative response rate	AUC	KS statistic	Error rate
Neural networks	92.843	**0.905**	**0.653**	**0.153**
Neural networks(boosting)	95.092	0.900	0.651	0.292
Neural networks(bagging)	**93.456**	0.903	0.647	0.163

Neural networks are experimented in the research. They are not deep learning methods but general neural networks which are applied in the research. Table 5 describe the results when neural networks are applied. In this case, neural networks without ensemble approach show the best performance like the above results.

Table 6. classification results by the combination of two classification methods

Combination	Top 10% cumulative response rate	AUC	KS statistic	Error rate
Regression, neural networks	92.638	0.904	0.653	0.153
Neural networks, decision tree	**98.978**	**0.912**	0.655	**0.142**
Regression, decision tree	97.751	0.908	0.647	**0.142**
Regression, NN, DT	98.160	0.910	**0.656**	0.144

Various combinations are describedin Table 6. We guessed that the complex combination would be the best model in this category before the experiment. However, the result of the experiment shows that the combination of neural networks and decision tree is better than the complex model by combination of regression, neural networks, and decision tree. We make an experiment for various situations and the results are displayed from Table 2 to 6. We choose the best model among them. We choose four models which are the best in each base classification, then we compare four selected models in Table 7. Ensemble with neural networks and decision tree shows the best performance among four selected models. Other classification method is better than ensemble with neural networks and decision tree on a specific measure. However, it is much better than others on overall performance.

Table 7. classification results among the best models

Combination	Top 10% cumulative response rate	AUC	KS statistic	Error rate
Decision tree	98.875	0.889	0.587	**0.141**
Regression	90.593	0.899	0.636	0.157
Neural networks	92.843	0.905	0.653	0.153
Neural networks, decision tree	**98.978**	**0.912**	**0.655**	0.142

IV. CONCLUSION

We have implemented several combinations with famous classification methods to Adult data set which is available on UCI Machine Learning Repository. We investigate the distributions of all variables and modify them by the logarithm transformation andthe group rare levels transformation. Many combination classifications are experimented. For each base classification method, a single classification method generally shows better performance than boosting

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

approach and bagging approach with the corresponding base classification. The combination of multiple base classification methods shows better than a single base classification. In the research, the combination of neural networks and decision tree is the best model for prediction of rich people in Adult data set.

We will develop the method to interpret the model for future research. the best model in the research is hard to understand since it includes neural networks. Also, the research procedure in the paper will be applied to other nation's census data since Adult data set is from US people and different people may have different characteristics.

REFERENCES

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal, "Data Mining", Morgan Kaufmann, 2016.
[2] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, 7(2), pp. 179-188, 1936.
[3] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees", Wadswoth & Brooks/Cole Advanced Books & Software, 1984.
[4] Warren McCulloch, and Pitts Walter, "A Logical Calculus of Ideas Immanent in Nervous Activity", Bulletin of Mathematical Biophysics, 5(4), pp. 115-133, 1943.
[5] Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoer, "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning", Automation and Remote Control, 25, pp. 821-837, 1964.
[6] Leo Breiman, "Bagging Predictors", Machine Learning, 24(2), pp. 123-140, 1996.
[7] Yoav Freund, and Robert E. Schapire, "A Decision-theoretic Generalization of On-line Learning and an Application to Boosting", Journal of Computer and System Sciences, 55(1), pp. 119-139, 1997.
[8] Leo Breiman, "Random Forests", Machine Learning 45(1), pp. 5-32, 2001.
[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The Elements of Statistical Learning", Springer, 2008.
[10] UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml/.
[11] R. Kohavi, "Scaling up the Accuracy of Naïve-Bayes Classifiers: A Decision-tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining", 1996.
[12] Wojtek J. Krzanowski, and David J. Hand, "ROC Curves for Continuous Data", Chapman and Hall/CRC, 2009.
[13] M. A. Stephens, "Test of Fit for the Logistic Distribution based on the Empirical Distribution Function", Biometrika, 66(3), pp. 591-595, 1979.