Education and Socio Economic Factors Impact on Earning for Pakistan - A Bigdata Analysis

Neelam Younas[1,2](✉), Zahid Asghar[1,2], Muhammad Qayyum[1], and Fazlullah Khan[3]

[1] Pakistan Institute of Development Econonics, Islamabad, Pakistan
qauidian2006@yahoo.com, g.zahid@gmail.com,
qayyum2494@gmail.com

[2] Department of Statistics, Quid-e-Azam University, Islamabad, Pakistan

[3] Department of Computer Science, Abdul Wali Khan University Mardan,
Mardan, Pakistan
fazlullah@awkum.edu.pk

Abstract. This paper give an insight on effect of education and socio economic factors on education on earning for Pakistan using data mining technique Regression tree and classification tree (CART). Labor force survey data used in this paper. Variables used as predictors in the study are Education, Gender, Status, Training, and Occupation, Location of working, Training, Experience, Age and Type of industry, where monthly income is used as an independent variable. In case of classification income is divided in Quintiles, which is used as a dependent variable for classification variable. Type of industry, education, age and occupation are found significant variables in both classification and regression tree. Regression trees shows that instead of education type of industry is the most important variable and sex and education are the least important variables. Classification tree also shows that Type of industry is the most significant variable which effects the earning of an individual, then age and occupation of an individual come and education is the least important variable where the rest of predictors play no role in earning of an individual.

Keywords: CART · Classification and regression tree · Pruning · Cross validation

1 Introduction

The distribution of the earnings is an important issue for the improving the socio economic condition of any country, especially when income distribution is skewed. To find the cause of difference in earnings of an individual or to find the determinants of earnings of individual whether personal characteristics play important role in effecting the earning of an individual or labor market characteristics. Once the factors effecting the earning of individual are known, then it is easy to improve life in that country. The predictor schooling used in Mincer earning function for Sweden and different cases when it yields misleading information and its assumptions about length of working life. It was found that the decline in rate to schooling from 1068 to 1981 in college

J. Ferreira and M. Alam (Eds.): Future 5V 2016, LNICST 185, pp. 215–223, 2017.
DOI: 10.1007/978-3-319-51207-5_22

education where return to high school is stable. There estimate suggests that impact of education on length of working life is an important topic for future research. Education has a causal effect on earnings (Bjorklund 2000).

The factors affecting the earnings of an individual and returns to education for Lahore district Pakistan for teaching and non-teaching staff in university, college and school using multiple linear regressions. The factors that significantly contributed to earning of all employees, university employees, college employees, and school employees were age, experience, occupation, gender, working hour, computer literacy, family background, and spouse education. Those who have passed SSC from private institute earn 8.7 more than those who have passed SSC from Government institute. Family background has positive and significant effect on earnings. Teaching staff earn more than non-teaching staff (Afzal 2011).

Earnings functions for industrial works in Punjab, to analyze the difference in earnings of individuals due to gender, marital status, regional location and other socio economic variables using linear single equation least squares regression analysis (Kapoor and Puri 1971). Parents effect the earning of a child potentially through genes and family environment by using variance component model to find the contribution of genetics, family and environments to the variance of the log earnings of white males around 50. The model is estimated through linear additive equation. The contribution of non-common environment is 46% for the log of earnings and 24% for the years of schooling. After making a lot of assumptions, they partition the remaining variance. Using more plausible estimates, the partitioning of the variance of the log of earning suggests 18 to 41% was due to genetics and 8 to 15% to common environment (Taubman 1976).

Decision tree is a flow-chart-like structure which is used for segmenting or stratifying the predictor space in to a number of regions or subsets, to make prediction for a given value, mean and mode of the training data set is used. The set of splitting rules used to segment the predictors space can be summarized in a tree, this approach is referred as Decision tree methods. The performance of tree based method and linear regression can be assess through test error where test error is estimated through cross validation or validation set. If the pictorial presentation of the model is required than we go for tree based methods. CART technique has used a lot in public health and finance but now-a-days used in economics.

We have used CART for finding the determinants of earning because of its interesting features. The purpose of using regression and classification tree (CART) is unlike simple regression its fit the model at each splitting node of the tree, where simple linear regression fit one model for the complete set of data. The Statistical earning function is given as follows, $\mathrm{Ln}\, y_i = f(s_i, x_i, z_i) + u_i$, ln y_i: is the log of earning, s_i: is schooling, x_i: is experience, z_i: Represents other factors affecting earning such as training of employees, gender or geographical region of individual, age, hours of work, type of industry the employees are working in u_i: is the disturbance term assumed to be normally distributed (Berndt 1991).

The data used in the study is that of Labor Force Survey 2012–13. We have used the information of only employed persons that is affecting earning of an individual i-e age, occupation, training, gender, experience, residence, educational level, marital status, and income. R-Programming have used for Classification and Regression Tree (CART) to estimate the determinants of earning function.

2 Empirical Results and Discussion

All individuals working in cooperative society, individual ownership, partnership and other, female their average log income is 8.325, so we make the prediction of $e^{8.325}$ i.e. 4125.737. Individuals working in cooperative society, individual ownership, partnership and other sectors but are females and having age less than 20 their average log income is 8.605. So we make prediction of $e^{8.606}$ i.e. 5458.885 but those whose age is greater than 20 their average log income is 9.0306. So the prediction is $e^{9.036}$ i.e. 8400. Those who are working in Government, private and public sector and having education below middle and no formal education and specifically working in private and public sector their mean log income is 9.246 so predicted as $e^{9.246}$ i.e. 10363 but working in the sector other than private and public and age is less than 36 their mean log income is 9.496, i.e. $e^{9.496}$. So prediction is 13306, having age greater than 36 their mean log income is $e^{9.825}$, *i.e.* 18490. So we conclude that the government employees earn more than other sectors employees. Females earn less than male. Employees having higher education and experience, earn most. Those females whose age is greater than 20 earn more than those, whose age is less than 20. This is depicted in Fig. 1.

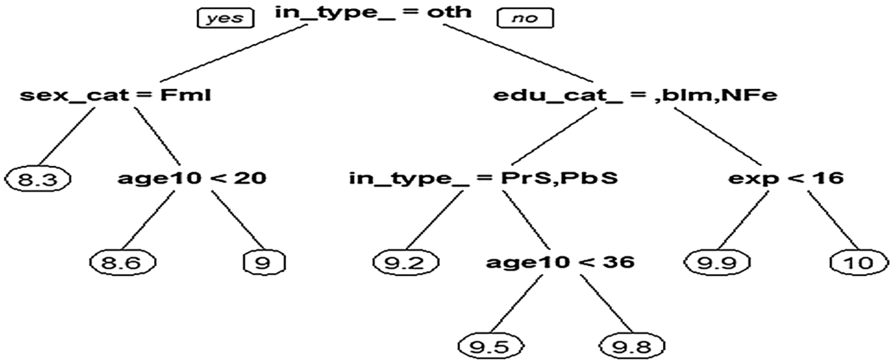

Fig. 1. Regression tree using complete data set of LFS

3 Making Prediction Through Fitted Model Using Testing Data

We have analyzed the data using fitted models as shown in Figs. 2 and 3. Cross validation graph shows the plot of size of the tree against the deviance. We choose that point where deviance approaches to minimum; here the minimum deviance is at size equal to 9. The pruned tree is shorter than the un pruned tree, the important variables are type of industry of an individual, sex and education. Pruned tree has five terminal

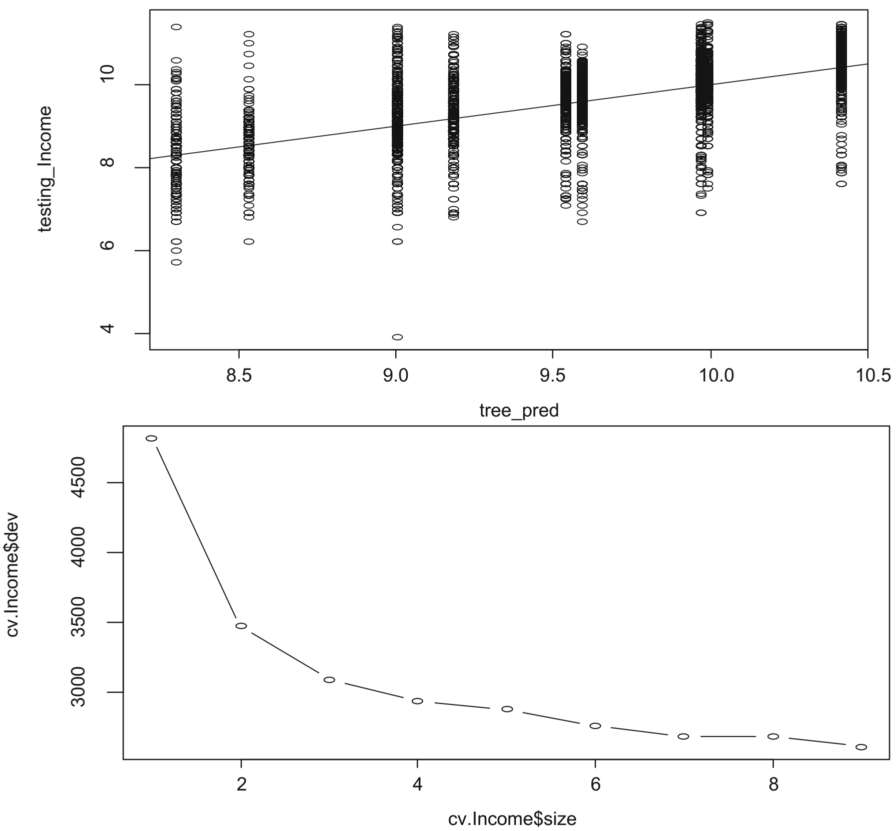

Fig. 2. Prediction made using testing data and cross validation

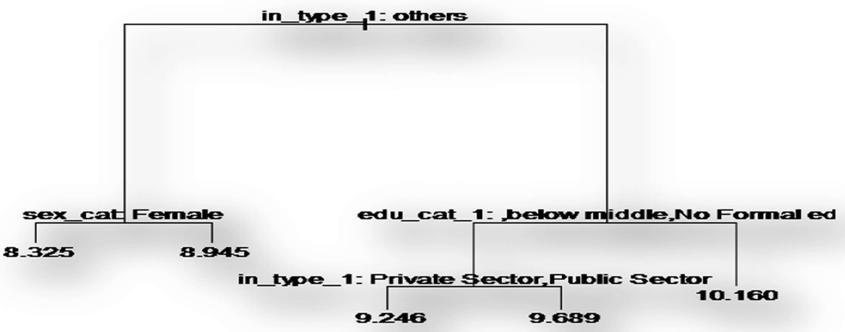
in_type_1: others
sex_cat: Female
8.325
8.945
edu_cat_1: ,below middle,No Formal ed
in_type_1: Private Sector,Public Sector
9.246
9.689
10.160

Fig. 3. Plot of prune tree

nodes and three internal nodes. Those individuals who are working in sectors other than government, private and public and are female their average log income is predicted as 8.3 i.e. 4023 and those who are males their average log income is 8.9 i.e. 7331. This shows males earn more than female if they are work in the same type of industry. Those who are working in government, private and public sector and education below middle or no formal education then specifically working in private and public sector their average log income is 9.2 i.e. 9897 and who are working in government or other sector their mean log income is 9.68 i.e. 15994. It shows that government employees with higher education earn more than employees of other sectors. Those who are not working in government, private and public sector in other words working in other sector and education greater than middle their mean log income is 10.160 i.e. 25648.

Figure 4 shows that the prediction using testing data through pruned model is quite good and the numerical measure used for calculating the error of the fitted model is MSE, which in this case is 36% and is increased a little for testing data.

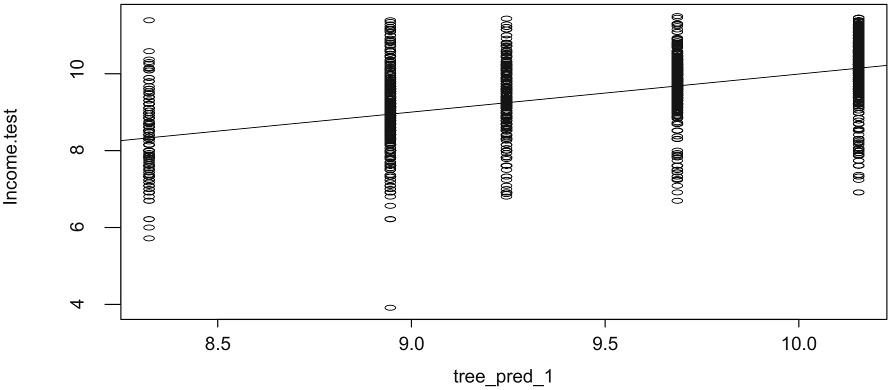

Fig. 4. Prediction through pruned tree using testing data

4 Classification Tree for Quintiles of Income

The classification tree depicted in Fig. 4, shows that those individuals who are working in government, private and public sectors, blue and pink collar workers and then working specifically in government sector are belongs to the Q1, the 1st quintile of income group. The range of 1st quintile is (0–16428), and those who are working in public, private or other sector also belong to Q1, the 1st quintile of income. Those who belongs to occupation category “white collar job” and “other”, age is less than 38.5 and education below middle belong to 1ist quintile of income. Those whose age is less than 38.5 but having education above middle and other also fall in 1st quintile of income.

Those whose age is greater than 38.5, education below middle and no formal education fall in 4th quintile (23273–29275) of income and those whose education category is other than below middle and no formal education fall in 5th quintile

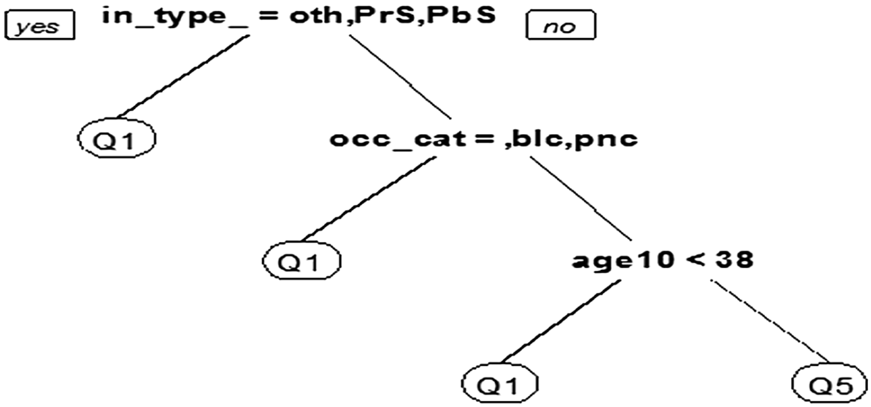

Fig. 5. Classification tree using R-part

(29276–46424) of income as shown in Fig. 5. So we conclude that individuals of "other sector" earn less than government, private and public sectors and belong to lower income group. Those who are working in government, private and public sectors, white collar workers, age greater than 38 and education below and above middle belong to the upper income group. Complete classification tree using training data in shown in Fig. 6.

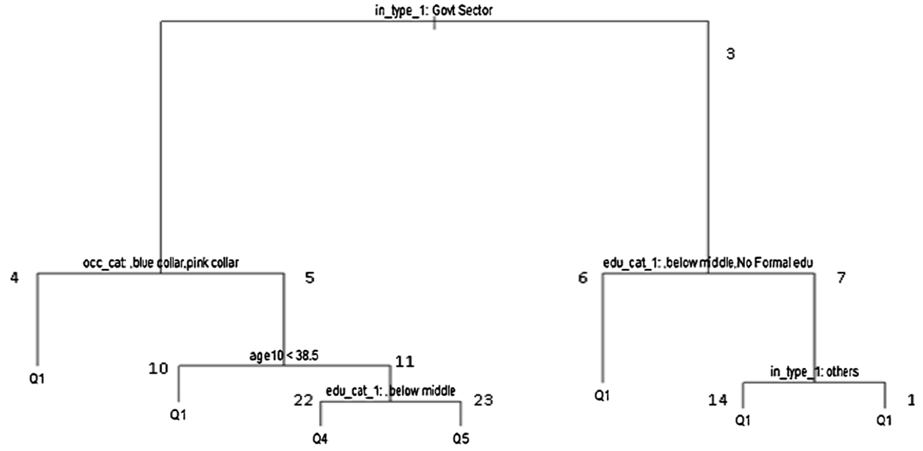

Fig. 6. Classification tree using training data

5 Cross Validation

Plot of the size of the tree against the misclassification shows the different size of tree against the misclassification but best point of pruning is 5, size of tree mean the number of leaves we have or the level to reach in pruning, when size of the tree is 5 the misclassification error is minimum. Figure 7 shows cross validation.

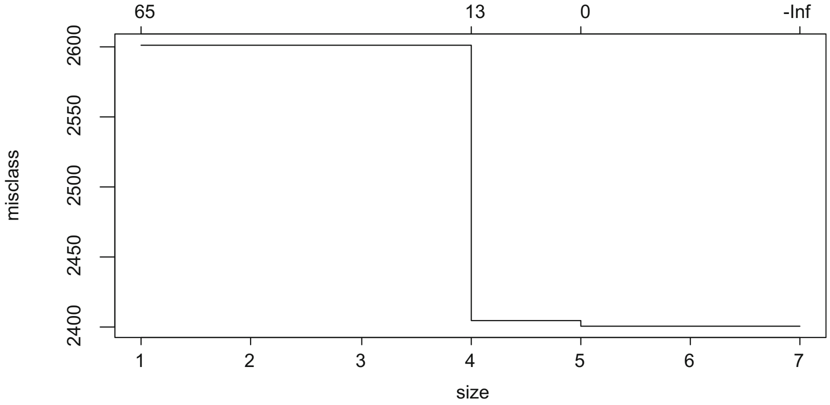

Fig. 7. Cross validation

6 Prediction Using Prune Tree

Left branch of the tree: Those individuals who are working in government sectors, blue and pink collar workers belong to the Q1, the 1st quintile of income group. The range of 1st quintile is (0–16428), and those who are white collar workers and their age is less than 38.5 also belongs to Q1 but those whose age is greater than 38.5 and having education middle or no formal education also belongs to Q4 but those whose education is above middle belongs to Q5 (5th quintile (29276–46424)). Individuals who are working in public, private or other sectors belong to Q1, the 1st quintile of income. Prediction by using pruned model for testing data/unseen data; Error is still 31% so the model is good fit for training and testing data (Fig. 8).

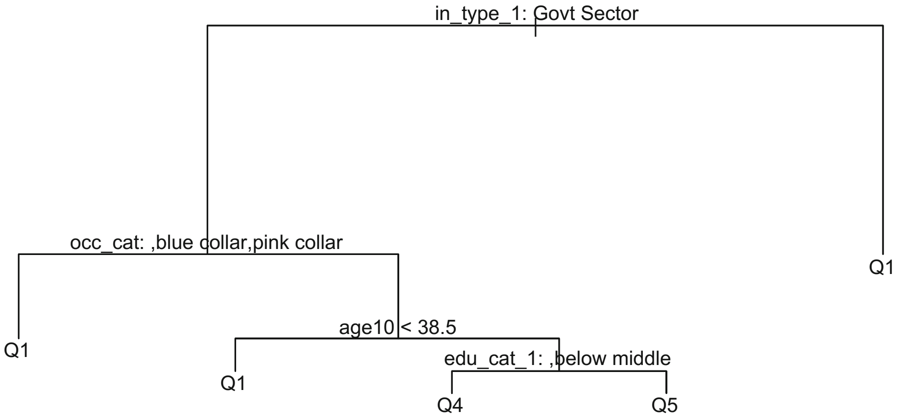

Fig. 8. Prune classification tree

7 Conclusion

In the final prune tree which is free from the problem of over fitting, three variables are significant, type of industry, sex and education. Here age has pruned and show that those individuals who are working in government, private and public sector and having higher education earn more than those whose education is lower than middle and those who are working in government sectors earn more than private and public sector. Female of Individual of cooperative society, individual ownership, partnership and other sectors earn less than male of these sectors. In case of classification, there are Quintiles of income, which is a qualitative variable. The classification tree made for Quintiles predicted that those individuals who are working in government, public and private sectors and are blue and pink collar workers then specifically in government sector, belongs to lower group income and if they are working in private and public sector also belong to lower group income. Those individual who are working in cooperative society, individual ownership, partnership and other sectors belong to lower income group if they are white collar workers and their age is greater than 38, having education above middle they belong to the highest income group but if individual have below middle education or no formal education and have age greater than 38, white collar worker and working in government, private or public sector belongs to income group Q4. So we conclude from classification tree that those individuals whose age is greater than 38, doing white collar job, having higher education and working in Government, private and public sector earn more. Type of industry an employee is working, occupation; education and age are important variable in the study.

References

Bjorklund, A., Kjellstrom, C.: Estimating the return to investments in education: how useful is the standard Mincer education? Econ. Educ. Rev. **21**, 195–210 (2000)

Metcalf, D.: The determinants of earnings changes: a regional analysis for the U.K., 1960–68. Int. Econ. Rev. **12**(2), 273–282 (1971)

Afzal, M.: Micro econometric analysis of private returns to education and determinants of earnings. Pak. Econ. Soc. Rev. **49**(1), 39–68 (2011)

Khan, S., Irfan, M.: Rates of returns to education and the determinants of earnings in Pakistan. Pak. Dev. Rev. **XXIV**(3&4), 671–683 (1985)

Tubman, P.: The determinants of earnings: genetics, family, and other environments; study of white male twins. Am. Econ. Assoc. **66**(5), 858–870 (1976)

Nasir, Z.: Determinants of earnings in Pakistan: findings from the labor force survey 1993–94. Pak. Dev. Rev. **37**(3), 251–274 (1998)

Kapoor, B.L., Puri, A.K.: The determinates of personal earnings: a study of industrial workers in Punjab. Econ. Educ. Rev. (1971)

Sutton, C.D.: Classification and regression trees, bagging, and boosting. In: Hand Book of Statistics, vol. 24 (2005)

Pakgohar, A., Tabrizi, R.S., Khalili, M., Esmaeili, A.: The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Comput. Sci. **3**, 764–769 (2010)

Lewis, R.J.: An introduction to classification and regression tree (CART) analysis. In: Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California (2000)

Berndt, E.R.: The Practice of Econometrics: Classic and Contemporary. Addison-Wesley, Boston (1991)

De'ath, G., Fabricius, K.E.: Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology **81**(11), 3178–3192 (2000)

Gordon, L.: Using classification and regression trees (CART) in SAS® enterprise miner TM for applications in public health. In: Data Mining and Text Analytics: 089-2013 (2013)

Horning, N.: Introduction to decision trees and random forests. Am. Mus. Nat. Hist. (2013)

James, G., Witten, D., Hastie, T.: An Introduction to Statistical Learning: With Applications in R. Taylor & Francis, Abingdon (2014)

Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**(3), 18–22 (2002)

Loh, W.Y.: Classification and regression trees. Wiley Interdisc. Rev.: Data Min. Knowl. Discov. **1**(1), 14–23 (2011)

Ohno-Machado, L., et al.: Decision trees and fuzzy logic: a comparison of models for the selection of measles vaccination strategies in Brazil. In: Proceedings of the AMIA Symposium. American Medical Informatics Association (2000)

Patel, H.D., et al.: Cost-effectiveness of a new rotavirus vaccination program in Pakistan: a decision tree model. Vaccine **31**(51), 6072–6078 (2013)

Rokach, L.: Data Mining with Decision Trees: Theory and Applications. World scientific, Singapore (2007)

Thakur, G.S., et al.: Understanding the applicability of linear & non-linear models using a case-based study. International Journal of Artificial Intelligence & Applications (IJAIA) **5**, 1–15 (2014)

Varian, H.R.: Big data: new tricks for econometrics. J. Econ. Perspect. **28**, 3–27 (2014)

Chang, Y.: Robustifying Regression and Classification Trees in the Presence of Irrelevant Variables. ProQuest, Ann Arbor (2008)

Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)

Zhang, D.: Advances in Machine Learning Applications in Software Engineering. IGI Global, Hershey (2006)