
Munich Personal RePEc Archive

Using decision tree classifier to predict income levels

Sisay Menji Bekena

30 July 2017

Online at https://mpra.ub.uni-muenchen.de/83406/
MPRA Paper No. 83406, posted 22 December 2017 05:10 UTC

Using decision tree classifier to predict income levels:

Sisay Menji Bekena

Abstract

In this study Random Forest Classifier machine learning algorithm is applied to predict income levels of individuals based on attributes including education, marital status, gender, occupation, country and others. Income levels are defined as a binary variable 0 for income <=50K/year and 1 for higher levels .The data is acquired from UCI Machine Learning Repository and includes 32,561 individuals data on 13 attributes based on 1994 census database. Random forest classifier is used since it gave better accuracy compared to decision tree classifier and naïve bayes classifier. The predictive accuracy of the model on test data is 85%. Important features prediction shows marital status, capital gain, education, age and hours per week are the top features which account for larger shares of the model accuracy. Using decision tree classifier also shows that these variables are the top 5 features in importance.

Motivation

Income inequality is one of the key issues governments are trying to solve. Reduced income disparity ensures balanced social development across different groups and improves the economic growth and political stability of a country. Governemnts in different countries are using different interventions to address income inequality, some are succedding while the others not. One of the key reasons behind failure is doing many things which results into reduced efficiency and lower results.

This study aims to conduct preliminary analysis that can be used to understand which factors are more important in improving individuals income. Such a study can help govenrments focus on a set of few (3-5) key areas that can significantly improve income levels of individuals. By focusing on few important areas governments can improve efficiency and achieve success at higher rates in reducing income inequality.

Dataset(s)

The data for the project was accessed from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Adult). The data is extracted by Barry Becker using 1994 census database.

The data set includes figures on 32,561 observations and 13 attributes for 42 countries. The target variable in the data set is income level which shows whether a person earns more than 50K (K denotes thousands, 50K equals 50,000) per year or not based on a set of different features. There are 12 features containing information on education, gender, nationality, marital status, occupation, work classification, gender, race, work hours, capital loss and capital gain.

Data Preparation and Cleaning

The following data preparation tasks are conducted to make the data suitable for running the machine learning model (decision tree classifier)

- **Converting categorical (text) values into dummy variables:** most of the variables are categorical (text) except capital gain, capital loss, hours per week, and years of education which are numeric. The categorical variables are transformed into dummy variables.
- **Dropping unnecessary columns and combining others:** a variable containing information on the sample weight of the individuals is dropped from since it is not required for the analysis. Capital gain and capital loss are merged into one column.
- **Checking for null values and preparing separate features and target data frames:** After doing all the above data cleaning steps, separates data frames for the feature and target data are generated.

Research Question(s)

Decision tree classifier is applied to the data to answer the following questions:

- **What is income level of an individual with certain attributes:** assessing whether an individual with certain attributes (age, education, sex, marital status and others) will earn higher or lower income levels. A higher income level is defined as an income level >50K/year while lower income level is defined as income <=50K/year. A function that predicts income level based on the fitted model and individual attribute is provided in the Ipython notebook.

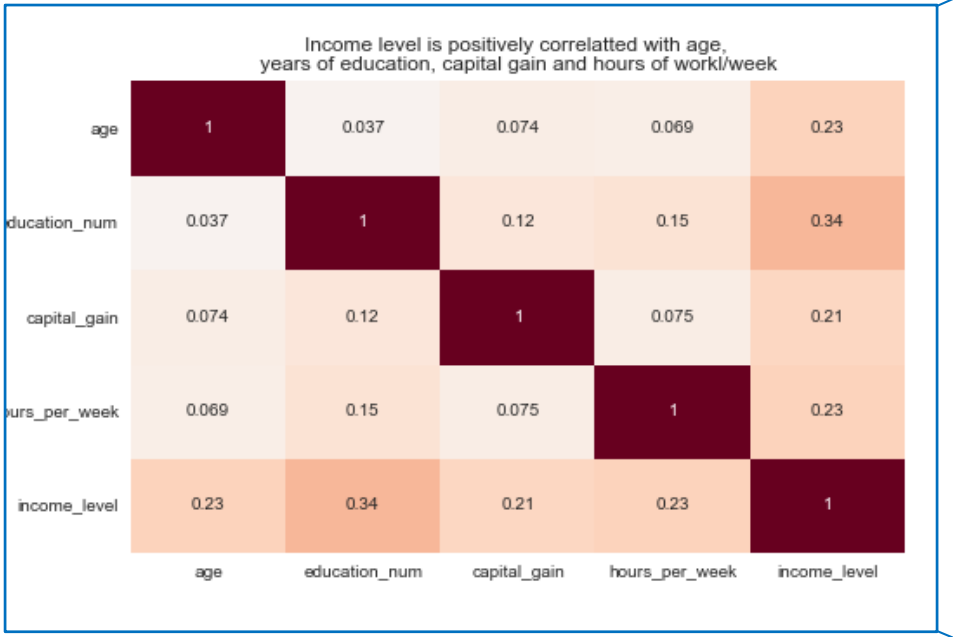
- **What are the key features determining income level:** identifying what are the top 5 features explaining much of the difference between low and high income levels. Determining the key features can help in policy formulation by identifying the few factors that can give most of the gains in income.

Methods

A supervised machine learning approach of **Random Forest Classifier** is used for the study. Random forest classifier is chosen due to two reasons. First since the outcome (target) variable is binary variable (income level >50K or not), using classification algorithms is better than regression algorithms. This is because the target having only values of 0 and 1, regression algorithms will perform less due to less variation in the target variable.

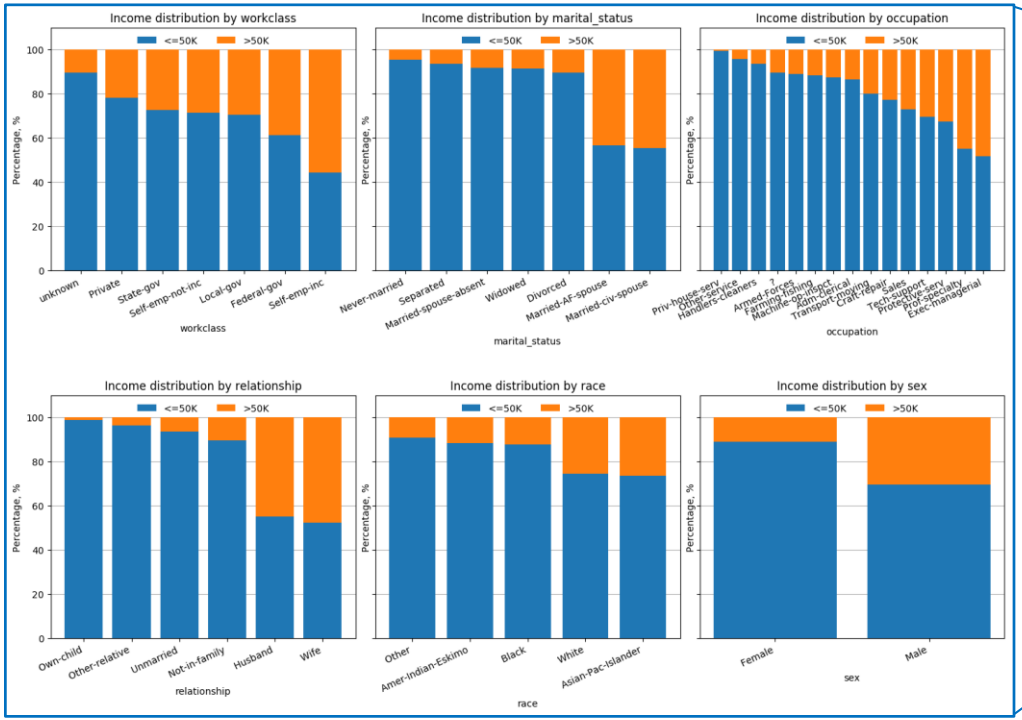
Secondly, random forest classifier is found to have better accuracy score compared to gaussian nainve baise classifier. Random forest and decisionTreeClassifier gave accuracy score of 85% while GaussianNB gave accuracy of 78%. Random forest is preferred to decision tree since using results from many decision treees will avoid the overfitting problem associated with using a decision tree classifier. The results from the model show that both random forest and decision tree gave similar results. This is shown by the fact that the top 7 features in importance are the same in the two models.

Education, capital, working more and age are found to have positive correlation with income



- Relatively education has the highest correlation +0.34 with income
- capital gain, age and hours worked per week are also positively correlated with income with a correlation coefficient of around 0.20.
- The variables are also positively correlated with each other with highest correlation observed between capital gain and education, and education and hours worked.

Plots of individuals with low and high incomes by different categorical variables shows the following results



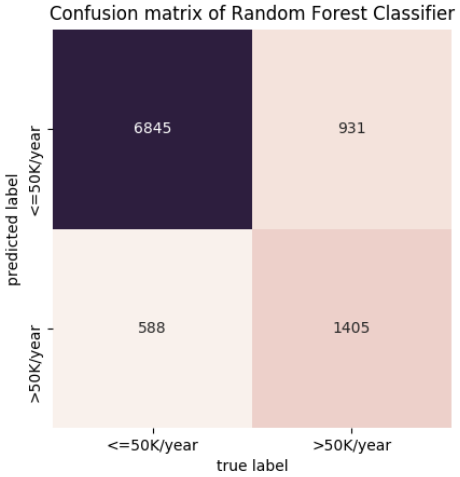
The following groups have higher share of individuals with higher income

- Employer: self employed and government employees
- Relationship: married people tend to have higher incomes
- Occupation: professionals, specialists, technology workers, and managers
- Race and gender: whites and asians earn higher compared to other races while men earn more compared to women

Random Forest Classifier (RFC) is chosen for the analysis

Since RFC model uses a number of decision tree classifiers to come up with a mean prediction, it is preferred to applying a single decision tree classifier

- Random forest classifier (RFC) is chosen because it has higher accuracy compared to GaussianNB (85% vs. 78%).
- RFC has similar accuracy score like decision tree classifier, but RFC is preferred since it reduces the overfitting tendency of decision tree classifier
- The model has a good accuracy on low income levels with only 7% of the values incorrectly identified as high income level but the performance is low on higher income levels with 40% of individuals with higher income predicted to have lower income (see confusion matrix on the right)
- The accuracy of the model is 85%. This should be seen in perspective since the model is not good in predicting higher income levels.



The top 5 important features account for 67% of the importance in the features[1]

Features	Importance	Cumulative sum
Age	23%	23%
Capital Gain	15%	38%
Education (# of years of schooling)	14%	52%
Hours Per Week	11%	63%
Marital Status Married	4%	67%
others	33%	100%
Total	100%	

Improvements in access to education, capital assess and employment opportunities can improve incomes significantly

- **Capital gain is accounts for ~15% of the variation**. This implies that improving access to capital is a key factor in improving income
- **The more educated a person is the more likely he/she will have higher incomes**. Improving access to education should be high priority for governments as it improves incomes.
- **Employment is also key:** people who work more will have higher incomes.

Note: feature importance is measured by the % decline in model accuracy caused by permuting a given variable.

The top 5 features identified using RFC are also in the top 5 features using decision tree classification

The top five features using RFC are also the top features using random forest classifier though now their total feature importance is ~70% compared to 67% for RFC.

Features	RFC Importance	RFC rank	RFC Cumulative	DTC Importance	DTC rank
Age	23%	1	23%	24%	1
Capital Gain	15%	2	38%	15%	2
Education (years of schooling)	14%	3	52%	13%	3
Hours Per Week	11%	4	63%	11%	4
Marital Status Married	4%	5	67%	6%	5
Others	33%		100%	31%	
Total	**100%**			**100%**	

Limitations

The analysis in this study has the following limitations.

- **Fitting data for 42 countries is difficult:** using cross sectional data will reduce the accuracy of any model applied since it is difficult to assume variables will have same impact in all countries. As an example the return to education in the United States and Philippines will be different and taking a signle figure for the different features across 42 countries will reduce the accuracy of the model.

- **Using the model for prediction is a bit difficult:** though the model will have relatively good accuracy using 1994 census data to predict about income levels today is difficult as things have changed much in the last 25 years. **Despite the above mentioned limitations, the model can be a good instrument in understanding which variables are key.**

Conclusions (1 of 2)

- This study used random forest classifier to predict income levels of an individual. Random forest classifier is chosen for the study because the target variable is categorical (binary - <=50K and >50K) and also because it has higher accuracy compared to naïve bayes classifier. Though the model accuracy is 85%, the model is weak in predicting high income individuals.

- The model is based on 1994 survey data. This makes predicting current income levels difficult since now income has increased in many countries compared to their levels in 1994 and economies have changed. **Despite this limitation, the model is useful in identifying the key factors that explain the difference between high and low income.**

Conclusions (2 of 2)

- Results from the fitted classifier model show that marital status, capital gain, education, age and work hours (employment) determine much of the difference between low and high income levels.
- Decision tree classifier gave same results to random forest classifier (this is partly because random forest is just applying a series of decision trees and taking the averages).

Acknowledgements

The data for the study assessed from University of California Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Adult). I would like to thank the University of California for providing such useful data sets for researchers for free. (I would also like to thank researchers and organizations who are sharing their data for others)

My work on this paper is based on the knowledge I gained from the edX course and other readings. I would like to thank data science instructors for the knowledge, guidance and materials provided to aid me (and other students) in their assignment.

References

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Scikit-learn documentation, http://scikit-learn.org/stable/documentation.html.