A Statistical Approach to Adult Census Income Data classification 1994

Zian MD Afique Amin, Khan Nasik Sami, Md Sariful Islam, Suhib Ahmad Abdulla

Department of Computer Science, Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia, Kuala Lumpur, Malaysia.
afiquezian@gmail.com nasiksami@gmail.com, mdsarif2435@gmail.com, suhibah@gmail.com

Abstract: In this study Gradient Boosting Classifier is used as the machine learning algorithm which is applied to predict income levels of individuals based on attributes including education, marital status, gender, occupation, country and others. The data for our study was accessed from the University of California Irvine (UCI) Machine Learning Repository. It was actually extracted by Barry Becker using the 1994 census database. The data set includes figures on 48,842 different records and 14 attributes for 42 nations. The 14 attributes consist of 8 categorical and 6 continuous attributes containing information on age, education, nationality, marital status, relationship status, occupation, work classification, gender, race , working hours per week, capital loss and capital gain. The binomial label in the data set is the income level which predicts whether a person earns more than 50 Thousand Dollars per year or not based on the given set of attributes. Gradient Boosting 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 86%. Which will eventually try to break the benchmark accuracy of existing works.

Keywords— Adult Census Data, Gradient boosting classifier, Classification, Machine learning

I. *Introduction:*

Over the last two decades, humans have grown a lot of dependence on data and information in society and with this advent growth, technologies have evolved for their storage, analysis, and processing on a huge scale. The fields of Data Mining and Machine Learning have not only exploited them for knowledge and discovery but also to explore certain hidden patterns and concepts which led to the prediction of future events, not easy to obtain. The problem of income inequality has been of great concern in recent years. Making the poor better off does not seem to be the sole criterion to be in the quest for eradicating this issue. People of the United States believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in society. This model actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary for improving an individual's income. Such an analysis helps to set focus on the important areas which significantly improve the income levels of individuals. Our chosen problem is a prediction problem.

II. *The Research Question and/or Hypothesis*

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 will be provided.

III. 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.

iv. Hypothesis:

Despite this limitation, using boosting algorithm the model will help in understanding how the income of a person varies depending on various factors such as the education-background, occupation, marital status, geography, age, number of working hours/week, etc.

V. The research objectives:

1. Import the data from the source and put it in the relevant platform. (Python/R)

2. Data Cleaning according to the needs.

3. Data Exploration, which involves analyzing each feature variable to check if the variables are significant for building the model.

4. Building a model, in which part we'll build a predictive model that will predict whether an individual earns above USD 50,000 or not.

5. Checking the accuracy of the model.

6. Load and evaluate the test data set in such a way that it does not have any null values or unnecessary predictor variables.

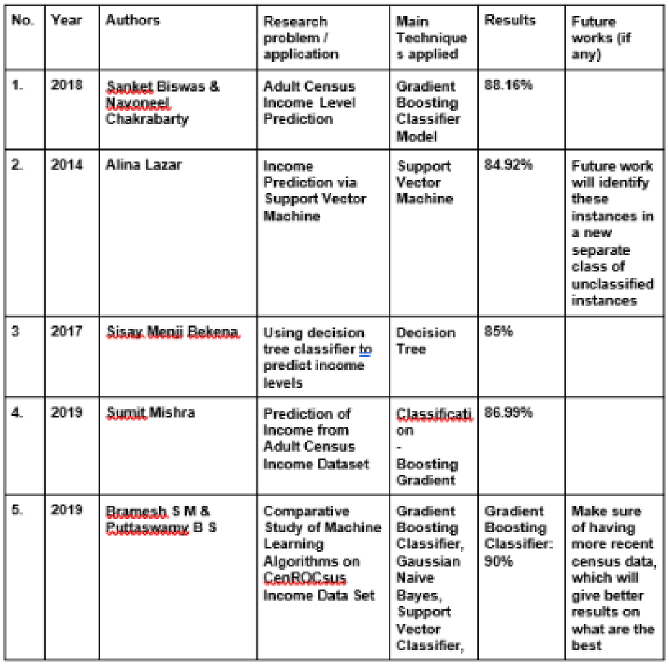
7. Validate the model.

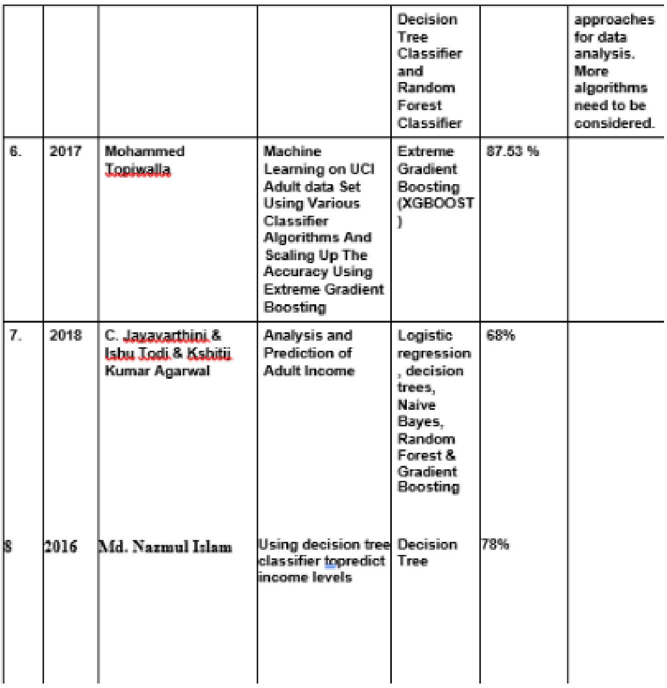
8. Show how the income of a person varies depending on various factors.

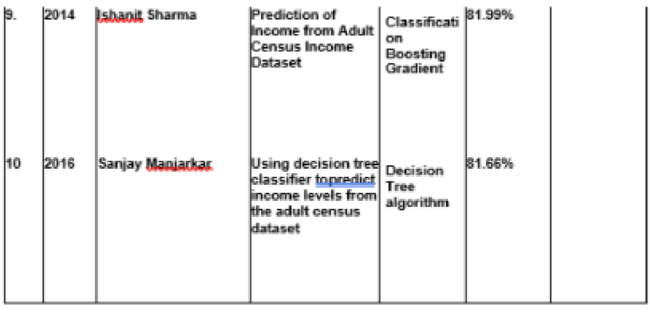
Vi. The Research Significances

This prediction will help us to get usable information, which can contribute to the business domain or how the income of a person varies depending on various factors such as the educational background, occupation, marital status, geography, age, number of working hours/week, etc. And 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.

VII. Related work







No.	Year	Authors	Research problem / application	Main Techniques applied	Results	Future works (if any)
1.	2018	Sanket Biswas & Navoneel Chakrabarty	Adult Census Income Level Prediction	Gradient Boosting Classifier Model	88.16%	
2.	2014	Alina Lazar	Income Prediction via Support Vector Machine	Support Vector Machine	84.92%	Future work will identify these instances in a new separate class of unclassified instances
3	2017	Sisay Menji Bekena	Using decision tree classifier to predict income levels	Decision Tree	85%	
4.	2019	Sumit Mishra	Prediction of Income from Adult Census Income Dataset	Classification - Boosting Gradient	86.99%	
5.	2019	Bramesh S M & Puttaswamy B S	Comparative Study of Machine Learning Algorithms on CenROCsus Income Data Set	Gradient Boosting Classifier, Gaussian Naive Bayes, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier	Gradient Boosting Classifier: 90%	Make sure of having more recent census data, which will give better results on what are the best approaches for data analysis. More algorithms need to be considered.
6.	2017	Mohammed Topiwalla	Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting	Extreme Gradient Boosting (XGBOOST)	87.53 %	
7.	2018	C. Jayavarthini & Ishu Todi & Kshitij Kumar Agarwal	Analysis and Prediction of Adult Income	Logistic regression, decision trees, Naive Bayes, Random Forest & Gradient Boosting	68%	
8	2016	Md. Nazmul Islam	Using decision tree classifier topredict income levels	Decision Tree	78%	
9.	2014	Ishanit Sharma	Prediction of Income from Adult Census Income Dataset	Classification Boosting Gradient	81.99%	
10	2016	Sanjay Manjarkar	Using decision tree classifier topredict income levels from the adult census dataset	Decision Tree algorithm	81.66%	

From this we can see that in recent years many people have tried to solve this income level prediction problem with this dataset. People tried to solve this using various methods and algorithms. Such as Gradient Boosting Classifier Model,Support Vector Machine,Decision Tree ,Gaussian Naive Bayes,Random Forest Classifier ,Extreme Gradient Boosting (XGBOOST) etc. The success rate of the prediction is varying from 85-90%.But we can see that the Gradient Boosting Classifier Model has been the most successful and consistent among them.

VIII Data Description

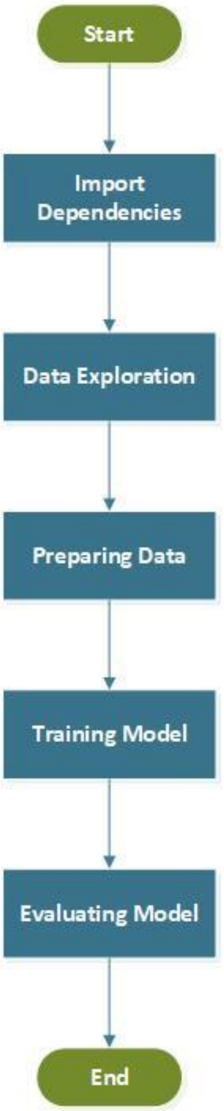
T

The data for our study was accessed from the University of CalifornIrvine (UCI) Machine Learning Repository. It was actually extracted by Barry Becker using the 1994 census database and given to the public website which is available ont the internet http://archive.ics.uci.edu/ml/datasets/Census+Income. This data set will help to understand how the income of a person varies depending on various factors such as the education background, occupation, marital status, geography, age, number of working hours/weeks, etc. The data set includes figures on 48,842 different records and 14 attributes for 42 nations.

IX. Experimental setup

The experimental setup explains the details of steps taken to figure out the answers of the research questions as mentioned previously. The general overview of data product is to apply the prediction methods for find out the winning rate of teams playing in the home ground or away. Therefore, the regression methods are necessary to search for the correct answers from the datasets

.



X. Algorithms

So, after evaluating all our predictor variables, it is finally time to perform Predictive analytics. In this stage, we'll build a predictive model that will predict whether an individual earns above USD 50,000 or not based on the predictor variables that we evaluated in the previous section.To build this model I've made use of the boosting algorithm since we have to classify an individual into either of the two classes, i.e:

Income level <= USD 50,000
Income level > USD 50,000

Gradient Boosting Classifier

The learning algorithm used to build the predictive model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Classifier. Unlike many ML models which focus on high quality prediction done by a single model, boosting algorithms seek to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors.

1) Boosting: It is an assembling technique, in which predictors (which are decision trees) are being constructed sequentially rather than independently.

2) Implementation: In GBC, in the sequence of predictors being constructed, at each and every sequence the error in the previous decision tree is corrected by the decision tree following it.

So, at each step, the GB Classifier, tends to fit the Training Data more and more. The Pseudo Code for Gradient Boosting

XI. Data Cleaning

The datasets to be used are first examined, and then to be cleaned as needed. The data cleaning stage is considered to be one of the most time-consuming tasks in Data Science. This stage includes removing NA values, getting rid of redundant variables and any inconsistencies in the data. We'll begin the data cleaning by checking if our data observations have any missing values.
The above code snippet indicates that 2399 sample cases have NA values. In order to fix this, let's look at the summary of all our variables and analyze which variables have the greatest number of null values. The reason why we must get rid of NA values is that they lead to wrongful predictions and hence decrease the accuracy of our model.

Attribute	Min	Q1	Median	Mean	Q3	Max
Age	17.00	22.00	36	40.39	58.00	90.00
Fnlwgt	12285	121804	177906	189543	232669	981628
Education	661	61	311	113	96	478
Educationnum	1	9	0	9.57	11	16
Maritalstatus	102	83	74	69	66	162
houseperwork	1	0	40	34.23	40	99
Capitalgain	0	0	73.87	0	0	4356
capitalloss	0	0	897.1	0	0	9999.0

XII. Data Exploration

Data Exploration involves analyzing each feature variable to check if the variables are significant for building the model.

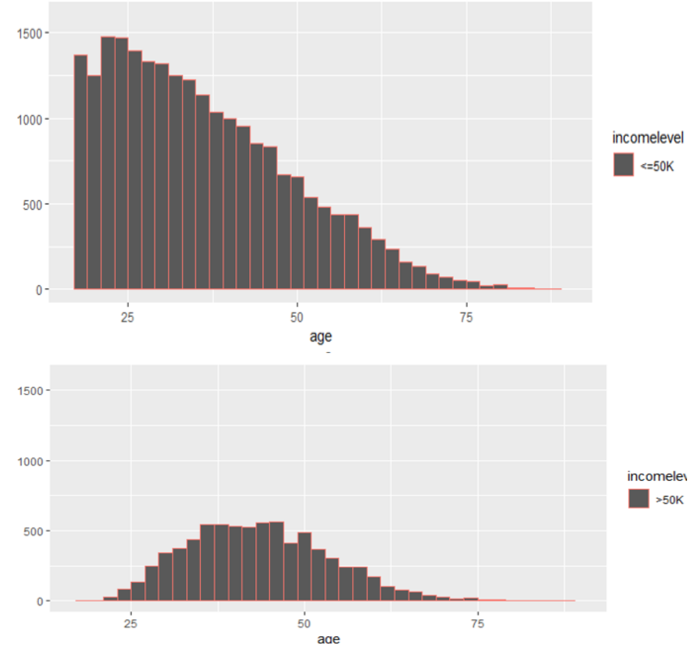


Fig: exploring data

XIII. Visualization Distribution of Data

Data Visualization has been done using Box and Whisker Plots of all continuous features to clearly understand the measures of their central tendencies shown in Fig 3, Fig 4, Fig 5, Fig 6, Fig 7 and Fig 8.

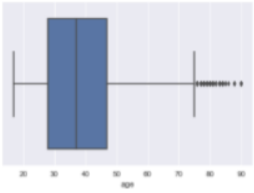


Fig 3. Box and Whisker for 'Age' attribute

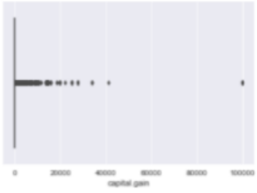


Fig 6. Box and Whisker for 'capital.gain' attribute

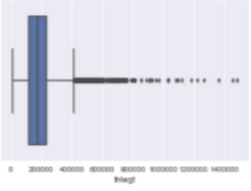


Fig 4. Box and Whisker for 'fnlwgt' attribute

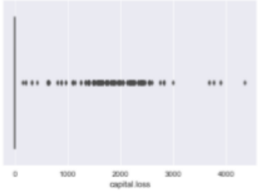


Fig 7. Box and Whisker for 'capital.loss' attribute

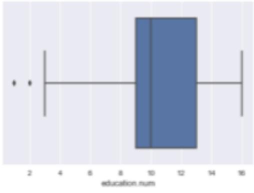


Fig 5. Box and Whisker for 'education.num' attribute

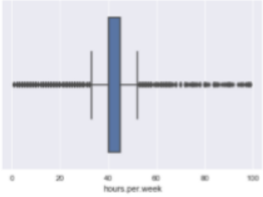
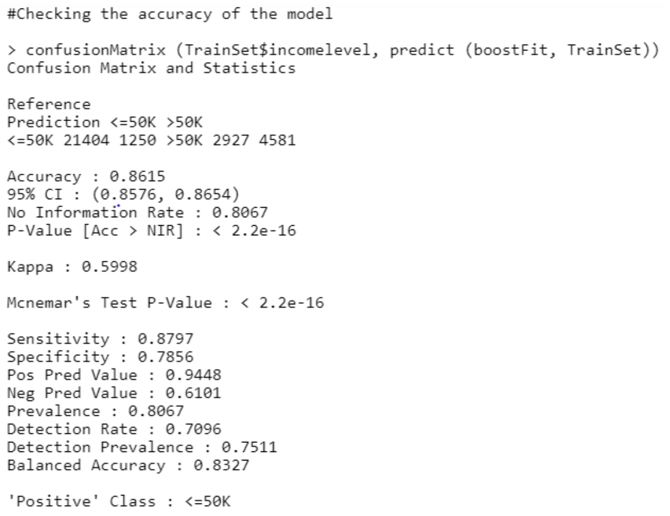


Fig 8. Box and Whisker for 'hours.per.week' attribute

XIV. Result Analysis

The result analysis is based on validation from datasets. Following are the result analysis below:



```
#Checking the accuracy of the model

> confusionMatrix (TrainSet$incomelevel, predict (boostFit, TrainSet))
Confusion Matrix and Statistics

Reference
Prediction <=50K >50K
<=50K 21404 1250 >50K 2927 4581

Accuracy : 0.8615
95% CI : (0.8576, 0.8654)
No Information Rate : 0.8067
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5998

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.8797
Specificity : 0.7856
Pos Pred Value : 0.9448
Neg Pred Value : 0.6101
Prevalence : 0.8067
Detection Rate : 0.7096
Detection Prevalence : 0.7511
Balanced Accuracy : 0.8327

'Positive' Class : <=50K
```

The output shows that our model calculates the income level of an individual with an accuracy of approximately 86%, which is a good number.

XV. Conclusion and Future Works

This paper proposed the application of Gradient Boosting Classifier on Adult Census Data. Finally, the Validation Accuracy, so obtained, 86% which is, by the best of our knowledge, has been one of the highest numeric accuracy achieved by Income Prediction Model so far. The classification error for the adult dataset is relatively large (~14%) because multiple similar instances have different decision values. The future scope of this work involves achieving an over-all better set of results by using hybrid models with inclusion of Machine Learning and Deep Learning together, or by applying many other advanced preprocessing techniques without further depletion in the accuracy. Future work will identify these instances in a new separate class of unclassified instances. Also the classification rate can be improved by using the kernel PCA method. This method consists in the classical PCA method applied to the kernel feature space. A speed up of the learning algorithm should result since most of the calculations are done at the kernel level.

References

[1] Vidya Chockalingam, Sejal Shah and Ronit Shaw: "Income Classification using Adult Census Data", https://cseweb.ucsd.edu/classes/wi17/cse258- a/reports/a120.pdf.
[2] Sisay Menji Bekena:"Using decision tree classifier to predict income levels", Munich Personal RePEc Archive 30th July, 2017
[3] Mohammed Topiwalla: "Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting", University of SP Jain School of Global Management.

[4] Alina Lazar: “Income Prediction via Support Vector Machine”, Inter- national Conference on Machine Learning and Applications - ICMLA 2004, 16-18 December 2004, Louisville, KY, USA.
[5] S.Deepajothi and Dr. S.Selvarajan: “A Comparative Study of Classifica- tion Techniques On Adult Data Set”, International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October- 2012.
[6] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: “Predicting if in- come exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”, https://cseweb.ucsd.edu/ jm- cauley/cse190/reports/sp15/048.pdf.
[7] Haojun Zhu: “Predicting Earning Potential us- ing the Adult Dataset”, https://rstudio-pubs-static.s3.amazonaws.com/235617 51e06fa6c43b47d1b6daca2523b2f9e4.html
[8] https://archive.ics.uci.edu/ml/datasets/Adult
[9] https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d >
[10] Trafalis, T.B., Santosa B., Richman, M.B.: Tornado Detection with Kernel-Based Classifiers From WSR-D88 Radar. Submitted to: Darema, F. (ed.) Dynamic Data Driven Application Systems, Kluwer, 2004.
[11] H. Nuncz, C. Angulo and A. Catala, ”Rule Extraction from Support Vector Machine”, Proccedings of ESANN’2002 – European Symposium on Artificial Neural Networks, Bruges (Belgium), pp. 107-112, 24-26 April 2002.
[12] G. M. Fung,, O. L. Mangasarian, and J. W. Shavlik, “Knowledge-based Nonlinear Kernel Classifiers”, Data Mining Institute Technical Report 03-02. Computer Sciences Department, University of Wisconsin, 2003.
[13] J. Wang and C. Zhang, “Support Vector Machines Based on Set Covering”, Proc. Of the 2nd International Conference on Information Technology for Application (ICITA), Harbin, China, January 2004.
[14] P. S. Bradley and O. L. Mangasarian, “Feature Selection via Concave Minimization and Support Vector Machines”. In Machine Learning Proceedings of the Fifteenth International Conference(ICML '98), J. Shavlik, editor, Morgan Kaufmann, San Francisco, California, 82-90, 1998.
[15] U.S. Census Bureau, United States Department of Commerce. Retrieved from http://www.census.gov/ on 11/16/03.