RESEARCH ARTICLE OPEN ACCESS

Comparative Study of Machine Learning Algorithms on Census Income Data Set

Bramesh S M*, Puttaswamy B S**
**(Department of Information Science & Engineering, P.E.S. College of Engineering, Mandya*
*** (Department of Information Science & Engineering, P.E.S. College of Engineering, Mandya*
Corresponding Author : Bramesh S M

ABSTRACT
Payment plans are a key tactical area for success and growth of a knowledge based industry and also optimum salary offer is essential to retain high performance employees. One of the challenges that industries face very often is finding such income facts, based on several information about a current employee or a future employee. Given the characteristics of a current employee or a future employee like his / her demographic profile along with other information such as performance level, qualification, etc., prediction of the salary class can be done by using many well-known machine learning algorithms. But unluckily, those details of employee of any industry are generally not presented in public for performance evaluation of machine learning algorithms. In this paper, this limitation is overcome to some extent by using a public database (UCI census data set) which has most of the attributes available for a segment of population for salary prediction. i.e., this paper aimed at examining and investigating five well-known supervised machine learning classifiers namely Gaussian Naive Bayes, Gradient Boosting Classifier, Support Vector Classifier, Random Forest Classifier and Decision Tree Classifier using the UCI census data set to find out the best classification algorithm out of above stated five well-known classifiers. It also aimed to determine the most effective classifier to be used in this area. Finally from the investigation we found that Gradient Boosting Classifier performed well when compared with the other four classifiers.

Keywords - Census Income dataset, Decision Tree Classifier, Gaussian Naive Bayes, Gradient Boosting Classifier, Random Forest Classifier, and Support Vector Classifier.

Date Of Submission: 25-07-2019 Date Of Acceptance: 06-08-2019

I. INTRODUCTION

With more emphasis on knowledge based industry, the payment forecasting is becoming a key strategic area for industries to ensure continuous growth and success. One of the problems which industries face till today is retaining high performing employees and also hire talented people from other industries. In both the cases, salary is a key significant aspect of tempting current as well as future employees. Hence a better salary offer is extremely important for retaining or attracting employees to any industry.

Human Resource (HR) managers have understood that several factors affect the salary expectation of an employee and only his / her past performance or performance in an interview is not the only determiner of his / her expected salary. Hence, to make a final offer to an employee, recruiters need to weigh several factors, including demographic as well as others. Although experienced HR managers drive this exercise in discussion with the relevant department level manager, it is always a tough decision.

Any type of automated decision making system would be helpful for these decision makers to come up with suitable salary recommendations. In this work, a public data set accessible from the University of California, Irvine (UCI) repository is used for investigating five machine learning algorithms, namely Gaussian Naive Bayes, Gradient Boosting Classifier, Support Vector Classifier, Random Forest Classifier and Decision Tree Classifier for prediction of salary and also measured their comparative performances. Even though the data used in this work is not directly related to salary prediction of employees within an industry, nevertheless it can be generalized to be used in the prior scenario as this too deals with binary salary class prediction of a sector of the population who work for multiple organizations.

This paper aimed at examining and investigating five machine learning algorithms, namely Gaussian Naive Bayes, Gradient Boosting Classifier, Support Vector Classifier, Random Forest Classifier and Decision Tree Classifier using the UCI census data set (University of California, 1994), and this work can definitely be considered as a beneficial effort towards understanding the usefulness of these algorithms for the real salary prediction problem. Although there are several limitations, nevertheless the outcomes can be used in real problem settings.

II. RELATED WORK

Several related research efforts have been conducted that employed census data by some classification algorithms. However, there is a need to evaluate and improve the performance of supervised learning in census data. Over the centuries, several techniques have been developed to deal with this size of data. Some of these techniques include multivariate regression analyses, as well as a total range of statistical methods [1].

Chockalingamet. et. al. [2] investigated the Adult Census Data to come up with crucial and exciting attributes of the data. By using a variety of machine learning models like Stepwise Logistic Regression, Logistic Regression, Naive Bayes, Extra Trees, Decision Trees, k-Nearest Neighbor, SVM, Gradient Boosting and six configurations of Activated Neural Network performed a predictive task of classification and also drew a relative analysis of their predictive performances.

Bekena [3] proposed a Random Forest Classifier to predict income levels of individuals based on various attributes of 1994 census database and they got 85% predictive accuracy on the test data.

Topiwalla [4] proposed approach that shows the correct flow of approaching a machine learning problem by demonstrating feature engineering, feature selection by using easy algorithms like Naive Bayes, Decision Tree, SVM, KNN and then gradually moving to more complex algorithms like Random Forest, XGBOOST, and Stacking of models.

Lazar [5] implemented Support Vector Machine and Principal Component Analysis methods to produce and assess income prediction data based on the present population survey provided by the U.S. Census Bureau.

Deepajothiet. al. [6] tried to replicate Decision Tree Induction, Bayesian Networks, Rule Based Learning and Lazy Classifier techniques for the Adult Dataset and presented a comparative analysis of the predictive performances.

Lemon et. al. [7] attempted to recognize the significant features in the data that could help to optimize the complexity of dissimilar machine learning models used in classification tasks.

Haojun Zhu [8] attempted Logistic Regression as the Statistical Modeling tool and 4 dissimilar machine learning techniques namely Classification and Regression Tree, Neural Network, Support Vector Machine, and Random Forest for predicting income levels.

It is also reported that researchers at the Ottawa University applied the method of decision trees to the Canadian census data in order to expose influences of bilingualism at the start of the last century [9] [10].

From the review we observed that the census dataset from UCI has been used in several cases, but only some with the intention of using it for employee salary prediction. In fact, only few works is focused on providing a benchmark of the existing research done in the comparative study of classifiers on predicting the range of income of a person from census data.

III. METHODOLOGY

The aim is to find out a classifier which will result in maximum accuracy in prediction of salary class (> 50 K, <= 50 K) based on the given set (or subset) of features. Therefore the purposes of this paper include:

- To apply Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier and Gradient Boosting Classifier on the public data set (UCI census)
- To compare prediction performance of above classifiers in terms of Accuracy, area under Receiver Operating Characteristics Curve (ROC), and F-measure.

A. The Dataset

The data for this study was truly mined by Barry Becker using the 1994 census data set and the data were accessed from the University of California Irvine (UCI) Machine Learning Repository [11].

Data set info in brief:

Total number of entries in the data set = 32561 entries

Total Data columns in the data set = 15 columns

Column	Entries	Null / Non-Null	Data type
Age	32561	non-null	int64
Work Class	32561	non-null	object
Final Weight	32561	non-null	int64
Education	32561	non-null	object
Education Number	32561	non-null	int64
Marital Status	32561	non-null	object
Occupation	32561	non-null	object
Relationship	32561	non-null	object
Race	32561	non-null	object
Sex	32561	non-null	object
Capital Gain	32561	non-null	int64
Capital Loss	32561	non-null	int64
Hours per Week	32561	non-null	int64
Country	32561	non-null	object
Income	32561	non-null	int32

Table 1: Column / Attribute details of the data set

The information above reveals that there are no missing values in the data set.

B. Exploratory Data Analysis and Data Processing

From the Exploratory Data Analysis we found that the data set has six continuous attributes, namely Final Weight, Age, Capital Gain, Education Number, Capital Loss, Hours per Week and nine categorical attributes, namely Education, Work Class, Marital Status, Relationship, Occupation, Race, Country, Sex and Income. The target variable is "Income", and it is a dependent variable. The other variables are independent. The income is divided into two classes: <= 50 K and > 50 K (Binary classification problem).

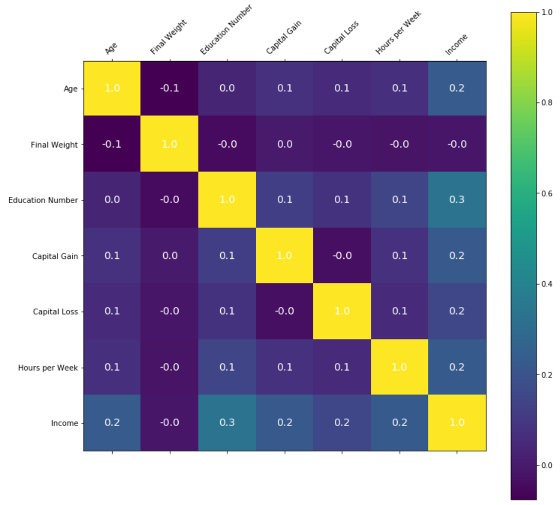

Fig. 1: Correlation matrix of the dataset

Taking a look at the correlation matrix above, it's clear that there is not a very high linear correlation between any of the continuous features / attributes and the target variable. Also, Final Weight has zero correlation with the output class and hence, we dropped this column from further analysis. Then we analyzed the categorical features / attributes using CountPlot (library function), which shows the counts of observations in each categorical bin using bars. Through analysis, we also found that there are some missing values in Country attribute. As they are very less, we have dropped these rows from further analysis. Then the whole data set has been mixed in a consistent way such that all the categories of dissimilar features remain included in Training Set and Validation Set.

Finally, the dataset is split into two sets, namely training and testing. Where 70% of the data is used for training purposes and the rest 30% of the data is used for testing purposes.

C. Applying Machine Learning

Here we have applied five algorithms to make the classification, namely Support Vector Classifier, Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier.

Python's Scikit-Learn Machine Learning Toolbox has been used for the Exploratory Data Analysis, Data Processing and Model Development. Python's Plotting Libraries like Matplotlib and Seaborn have been used for the data Visualizations.

D. Analyzing Results

After building the model, the most significant query that arises is how decent is the built model? So, assessing the built model is the most vital task which describes how good the model predictions are.

Accuracy - is the best natural performance measure and it is simply a fraction of properly predicted observation to the whole observations.

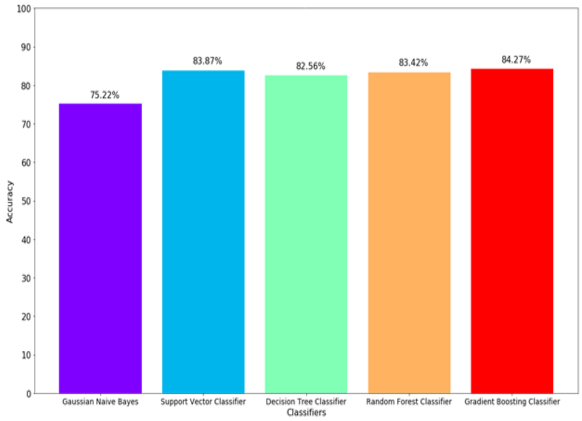

Fig. 2: Accuracy Plot of all classifiers

As it can be seen from the Fig. 2, the Gradient Boosting Classifier had the best accuracy when compared with Support Vector Classifier, Gaussian Naive Bayes, Random Forest Classifier and Decision Tree Classifier.

F1 score - is the weighted average of Precision and Recall.

$$F1\ score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \qquad [12]$$

Classifier	F1 score
Gaussian Naive Bayes	0.64
Support Vector Classifier	0.62
Decision Tree Classifier	0.62
Random Forest Classifier	0.64
Gradient Boosting Classifier	0.65

Table 2: F1 score of all classifiers

As it can be seen from the table above, the Gradient Boosting Classifier had the best F1 Score when compared with Support Vector Classifier, Gaussian Naive Bayes, Random Forest Classifier and Decision Tree Classifier.

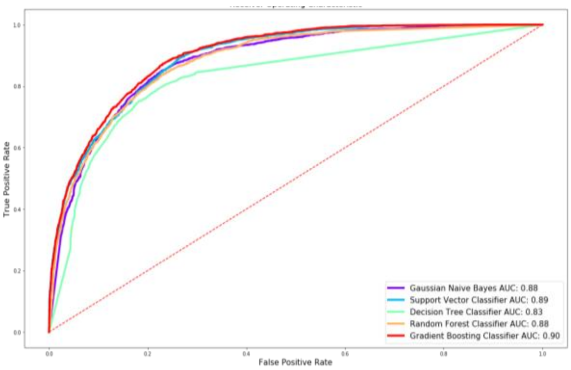

Fig**. 3:** Receiver Operating Characteristics Curve

The above figure shows the ROC for the predictions for income more than $50K. Again, Gradient Boosting Classifier has the maximum Area under curve with a value of 0.90.

IV. CONCLUSION

This paper aimed to examine and investigate five well-known supervised machine learning classifiers namely Gradient Boosting Classifier, Gaussian Naive Bayes, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier using the UCI census dataset to find out a classification algorithm which will result in maximum accuracy in prediction of salary class. It also aimed to determine the most effective classifier to be used in this area.

Gradient Boosting Classifier was considered to be the best classifier, since it had the highest ROC index with 0.90. It also had the highest accuracy and the lowest misclassification rate. There are lots of areas that can be carried out in the future. One of the main drawbacks of this study was that the data used in this study was not the recent census data. As a result, it is highly recommended to find more recent census data in order to make the models more suitable for today's populations. Another area of the future work is to investigate different classifiers for predicting the annual income.

REFERENCES

[1]. Sumathi, S., and Sivanandam, S. (2006). Introduction to Data Mining and its Applications.Springer-VerlagBerlin eidelberg. doi:10.1007/978-3-540-34351-6.

[2]. Vidya Chockalingam, Sejal Shah and Ronit Shaw: "Income Classification using Adult Census Data".

[3]. https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a120.pdf.

[4]. Sisay Menji Bekena: "Using decision tree classifier to predict income levels", Munich Personal RePEc Archive 30th July, 2017.

[5]. 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.

[6]. Alina Lazar: "Income Prediction via Support Vector Machine", International Conference on Machine Learning and Applications - ICMLA 2004, 16-18 December 2004, Louisville, KY, USA.

[7]. S.Deepajothi and Dr. S.Selvarajan: "A Comparative Study of Classification Techniques On Adult Data Set", International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October- 2012.

[8]. 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/jmcauley/cse190/reports/sp15/048.pdf.

[9]. Haojun Zhu: "Predicting Earning Potential using the Adult Dataset", https://rstudio-pubs-static.s3.amazonaws.com/23561751e06fa6c43b47d1b6daca2523b2f9e4.html

[10]. Hassani, H., Saporta, G., & Silva, E. (2014). DATA MINING AND OFFICIAL STATISTICS: The Past, the Present and the Future. The journal of big data, 2(1), 34-43. doi:10.1089/big.2013.0038.

[11]. Drummond, C., Matwinm, S., & Gaffield, C. (2000). Inferring and revising theories with confidence: data mining the 1901 Canadian census. Journal of Machine Learning Research, 1-48. doi:10.1080/08839510500313711.

[12]. https://archive.ics.uci.edu/ml/datasets/Adult

[13]. https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

Bramesh S M "Comparative Study of Machine Learning Algorithms on Census Income Data Set" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.08, 2019, pp. 78-81