Grenze International Journal of Engineering and Technology, July Issue

Census Employee Salary Prediction using Supervised Machine Learning

Raghawendra Naik[1] and Pavan N.Kunchur[2]
[1]Department of Computer Science and Engineering, KLS Gogte Institute of Technology,Belagavi
raghawendranaik@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, KLS Gogte Institute of Technology,Belagavi
pnkunchur@git.edu

***Abstract*—Payment plans are a key tactical area for fulfillment and growth of knowledge based industry and also optimum salary offer is essential to retain high performance employees. One of the challenges that industries face fairly 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 alongside other information like performance level, qualification, etc., prediction of the salary class are often done by using many well-known machine learning algorithms. But unluckily, those details of employee of any industry are generally not presented publicly for performance evaluation of machine learning algorithms. In this paper, this limitation is overcome to some extent by employing 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 three well-known supervised machine learning classifiers namely Gaussian Naive Bayes, *KNN(K-Nearest Neighbors)* and Decision Tree Classifier using the UCI census data set to find out the best classification algorithm out of above stated three 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 *KNN(K-Nearest Neighbors)* Classifier performed well in comparison with the opposite two classifiers.**

***Index Terms*— Census Income dataset, Decision Tree Classifier, Gaussian Naive Bayes, KNN(K-Nearest Neighbors).**

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

Grenze ID: 01.GIJET.6.2.503_2

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 three machine learning algorithms, namely Gaussian Naive Bayes, , KNN(K-Nearest Neighbors) 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 three machine learning algorithms, namely Gaussian Naive Bayes, KNN(K-Nearest Neighbors) 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. LITREATURE SURVEY

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, KNN(K-Nearest Neighbors) Classifier *and* Decision Tree 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).

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 [16].
- Data set info in brief:
- Total number of entries in the data set = 32561 entries
- Total Data columns in the data set = 15 columns

TABLE I: COLUMN / ATTRIBUTE DETAILS OF THE DATA SET

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

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).

Taking a look at the correlation matrix , 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 three algorithms to make the classification, namely Naive Bayes Classifier**,** KNN(K-Nearest Neighbors) Classifier and Decision Tree 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. the KNN(K-Nearest Neighbors) had the best accuracy when compared with Gaussian Naive Bayes Decision Tree Classifier.

IV. SYSTEM ARCHITECTURE

The proposes Prediction engine has been tested on to decide the salary for jobs in UK to improve the experience of peoples searching for jobs, and help employers and jobseekers to figure out the latest market worth for different positions. The dataset contains age, work class, fnlwgt, education, marital_status, occupation, relationship, race, sex,capital_gain,capital_loss,hours_per_week,native_country. Based on the above mentioned features our Prediction engine will predict salary.

A. Description of the Dataset:

The dataset used in this work is collected and uploaded to www.kaggle.com by Chet Lemon, Chris Zelazo from where we've downloaded this dataset. It contains 48,842 data points; a snapshot of the data set is given in Fig. 1. We have divided our dataset in two parts- train dataset and test dataset and the divide ratio is 1/3rd in test set and 2/3rd in train set.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	age	workclass	fnlwgt	education	education_	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income_level	
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K	
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K	
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K	
5	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K	
6	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K	
7	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K	
8	49	Private	160187	9th	5	Married-spouse-abs	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K	
9	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K	
10	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K	
11	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K	
12	37	Private	280464	Some-colleg	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K	
13	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac	Male	0	0	40	India	>50K	
14	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K	
15	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K	
16	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac	Male	0	0	40	?	>50K	
17	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indi	Male	0	0	45	Mexico	<=50K	
18	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K	
19	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K	
20	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K	
21	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K	
22	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K	
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K	
24	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K	
25	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K	

census_income_dataset

Fig 1: A Snapshot of the Dataset

Non target attributes

1. **age:** the age of an individual .
2. **workclass:** a general term to represent the employment status of an individual
3. **fnlwgt:** final weight. In other words, this is the number of people the census believes the entry represents..
4. **education:** the highest level of education achieved by an individual
5. **education_num:** the highest level of education achieved in numerical form.

6. **marital_status:** marital status of an individual.
7. **occupation:** the general type of occupation of an individual
8. **relationship:** represents what this individual is relative to others.
9. **race:** Descriptions of an individual's race.
10. **sex:** the biological sex of the individual.
11. **capital_gain:** capital gains for an individual.
12. **capital_loss:** capital loss for an individual.
13. **Hours_per_week:** the hours an individual has reported to work per week.
14. **native_country:** country of origin for an individual.
15. **the label:** whether or not an individual makes more than $50,000 annually.

Target attributes
Salary Raw – it is the amount which will be paid for the job.
The next step for preprocessing of raw data set is data analysis.

V. Results and Analysis

In this Proposed system, we have applied Three machine algorithms such as KNN(K-Nearest Neighbors) Classifier, Gaussian Naive Bayes Classifier, and Decision Tree Classifier to a Census dataset which is Trained and Tested. We found the result with accuracy of the salary which is less than or equal to 50K for three algorithms where the KNN has the highest accuracy compared to other algorithms because this dataset is the classified dataset.

TABLE II: Machine learning Algorithms with Accuracy

Algorithm	Accuracy
KNN(K-Nearest Neighbors) Classifier	87.81
Decision Tree Classifier	83.10
Gaussian Naive Bayes Classifier	79.62

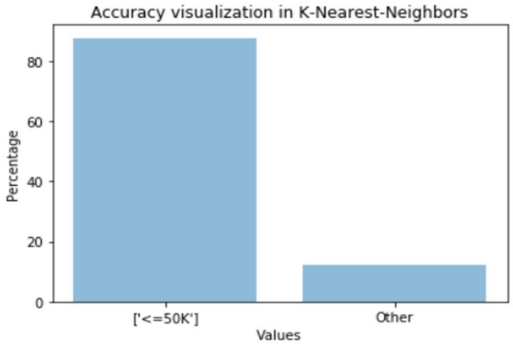

Fig 2: Accuracy visualization in KNN

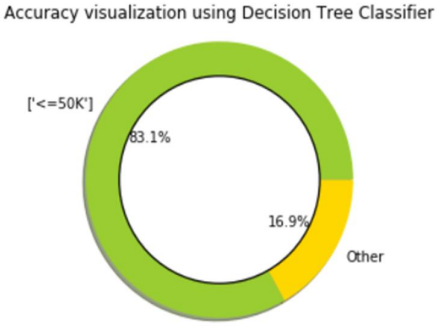

Fig 3: Accuracy visualization in Decision Tree Classifier

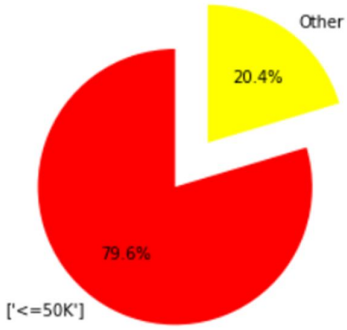

Fig 4: Accuracy visualization in Naïve – bayes

VI. Conclusion

This paper aimed to examine and investigate three well-known supervised machine learning classifiers namely KNN(K-Nearest Neighbors) Classifier, Gaussian Naive Bayes, and Decision Tree Classifier using the UCI census dataset to seek out a classification algorithm which can end in maximum accuracy in prediction of salary class. It also aimed to determine the most effective classifier to be used in this area.
KNN (K-Nearest Neighbors) 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/