Income Classification Using Logistic Regression

Huzaifa, Pratik Tabhane, Gopichand Gudduru, Pothagani Keerthi, Nidhi Lal

Dept. of Computer Science and Engineering

IIIT Nagpur, India

huzaifa060200@gmail.com,

tabhanepratik02@gmail.com

gopichandgudduru@gmail.com

keerthipothagani@gmail.com

nidhi.2592@gmail.com

Abstract – The prominent inequality of wealth and income is a huge concern especially in India. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improve the economic stability of a nation. Governments in different countries have been trying their best to address this problem and provide an optimal solution. This project report aims to show the usage of machine learning and data mining techniques to provide a solution to the income inequality problem.Our work aims to predict whether the income of U.S. population exceeds $50K/yr or not based on census data provided by Census bureau database, while considering different factors such as age, work class, gender, marital status, education, race, occupation etc. using exploratory analysis and classification algorithms. The dataset contains 48843 records. The dataset encourages to draw valuable insights and conclusions. The conclusions drawn might help in delivering wiser decisions. In addition to it, suggestions could be given based on the predictions to students who are in need to pursue higher education and people who are spending less time in the workplace.

Keywords **– machine learning, data mining, income equality, classification, logistic regression.**

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[1]. 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[2]. This model actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income[3]. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals[4]. This paper has been structured as an introduction, literature review, proposed methodology, training the model, implementation details, results and conclusion[5].

II. RELATED WORK

Certain efforts using machine learning models have been made in the past by researchers for predicting income levels.

Chockalingam et. al. [6] explored and analysed the Adult Dataset and used several Machine Learning Models like Logistic Regression, Stepwise Logistic Regression, Naive Bayes, Decision Trees, Extra Trees, k-Nearest Neighbor, SVM, Gradient Boosting and 6 configurations of Activated Neural Network. They also drew a comparative analysis of their predictive performances.Bekena [7] implemented the Random Forest Classifier algorithm to predict income levels of individuals.Topiwalla [8] made the usage of complex algorithms like XGBOOST, Random Forest and stacking of models for prediction tasks including Logistic Stack on XGBOOST and SVM Stack on Logistic for scaling up the accuracy.

Lazar [9] implemented Principal Component Analysis (PCA) and Support Vector Machine methods to generate and evaluate income prediction data based on the Current Population Survey provided by the U.S. Census Bureau. Deepajothi et. al. [10] tried to replicate Bayesian Networks, Decision Tree Induction, Lazy Classifier and Rule Based Learning Techniques for the Adult Dataset and presented a comparative analysis of the predictive performances.Lemonet. al.[11]

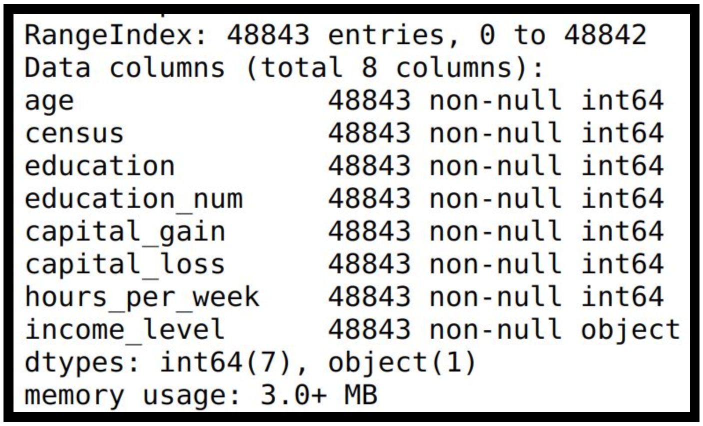
attempted to identify the important features in the data that could help to optimize the complexity of different machine learning models used in classification tasks.Haojun Zhu [12] attempted Logistic Regression as the Statistical Modelling Tool and 4 different Machine Learning Techniques, Neural Network, Classification and Regres-sion Tree, Random Forest, and Support Vector Machine for predicting Income Levels.

III. PROPOSED WORK

A. The Dataset

The dataset used in this project has 48,843 records and a binomial label indicating a salary of <=50K or >50K USD. 76% of the records in the dataset have a class label of <=50K. There are 8 attributes consisting of one categorical and seven continuous attributes (Fig. 1).

The employment class describes the type of employer such as self-employed or federal and occupation describes the employment type such as farming, clerical or managerial. Education contains the highest level of education attained such as high school or doctorate. The relationship attribute has categories such as unmarried or husband and marital status has categories such as married or separated. The other nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (numeric representation of the education attribute), capital gain and loss, and a weight attribute which is a demographic score assigned to an individual based on information such as state of residence and type of employment. People with similar demographic characteristics should have similar weights.



```
RangeIndex: 48843 entries, 0 to 48842
Data columns (total 8 columns):
age                 48843 non-null int64
census              48843 non-null int64
education           48843 non-null int64
education_num       48843 non-null int64
capital_gain        48843 non-null int64
capital_loss        48843 non-null int64
hours_per_week      48843 non-null int64
income_level        48843 non-null object
dtypes: int64(7), object(1)
memory usage: 3.0+ MB
```

Fig. 1showing the various attributes,

their types and their quantity used in the dataset.

B. Nominal Attributes

- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. 69.4% values are Private. 6% of values are unknown (‘?’)

- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. 32% are high school graduates, 22% went to some college and 16.5% have a bachelor’s degree

- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. 46% are married to civilian, 33% are never married, 14% are divorced

- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspectot, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. 6% values are unknown, evenly distributed except armed forces (0.03%) and private house servant (0.5%)

- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. 40% are husbands, 26% are not in a family

- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. 86% are White and 10% are Black with negligible proportions of others

- sex: Female, Male. 67% are male and 33% female

- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador,Taiwan, Haiti, Colombia, Hungary, Guatemala, Nicaragua, Scotland, Trinidad and Tobago, Peru, Hong, Holland-Netherlands. 42 categories, 90% are from United States, 2% values are unknown.

C. Label

Income: <=50K (76%), >50K (24%)

D. Data Visualization

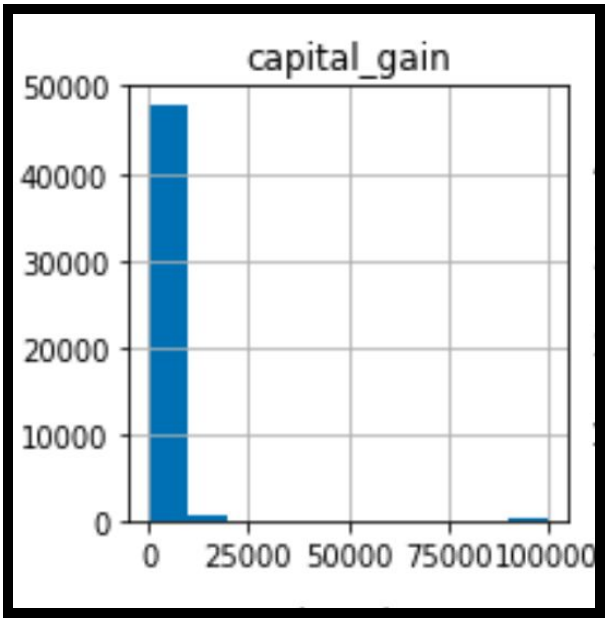


Fig. 2 Plot between Capital Gain and Income Level

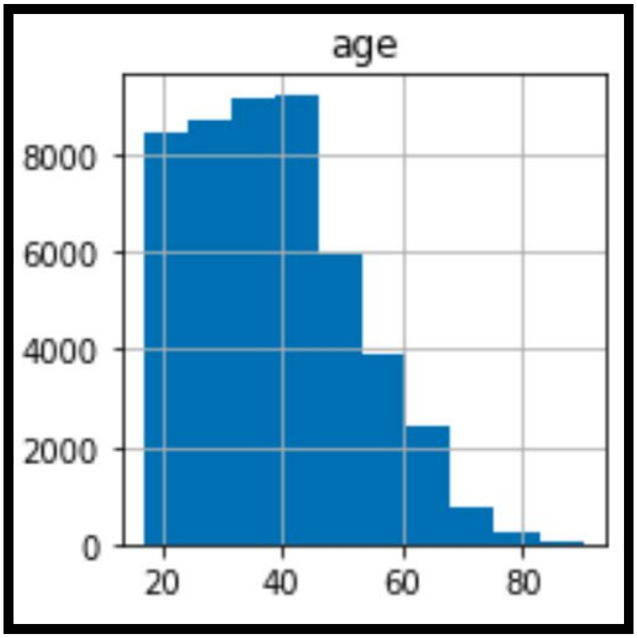


Fig. 3 Plot between Present Age and Income Level

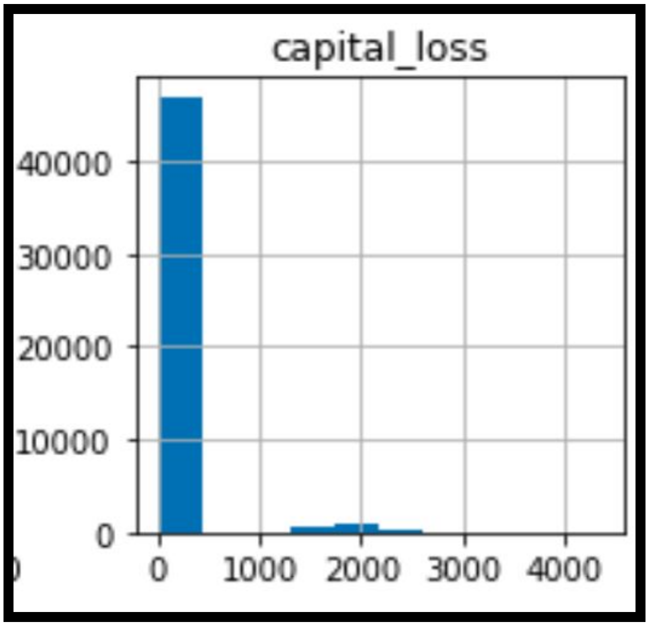


Fig. 4 Plot between Capital Loss and Income Level

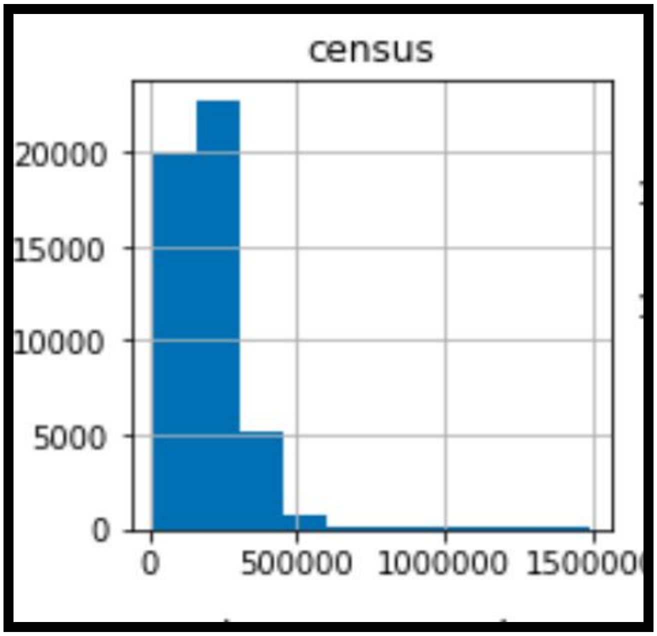


Fig. 5 Plot between Census and Income Level

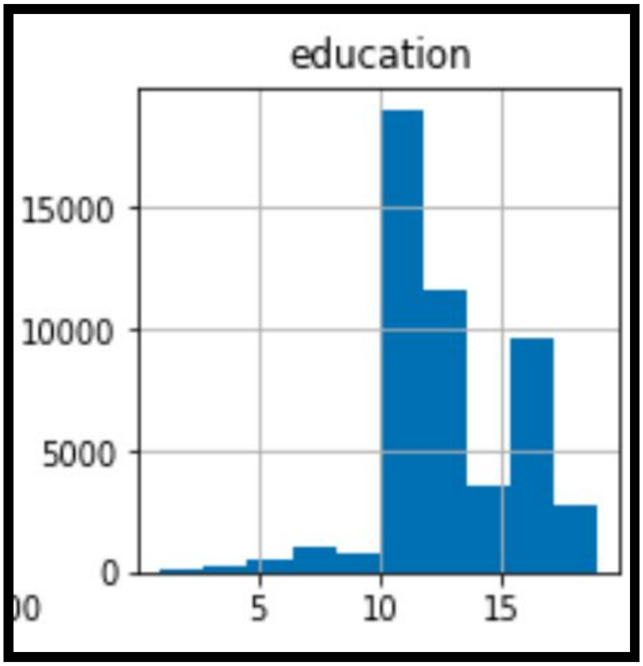


Fig. 6 Plot between Current Education Level and Income Level

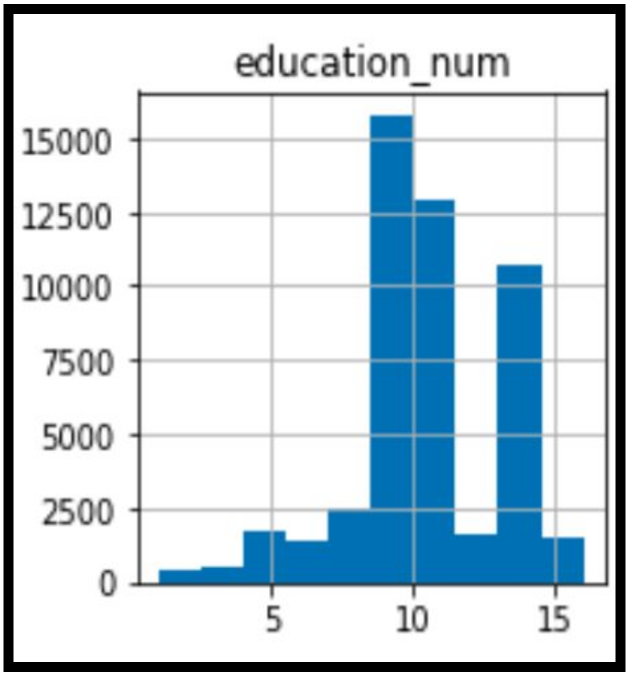


Fig. 7 Plot between Educational Qualifications and Income Level

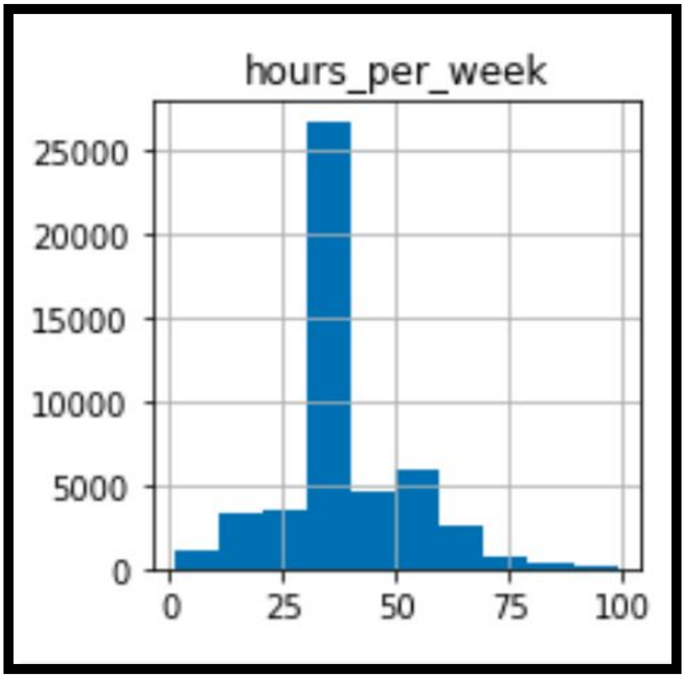


Fig. 8 Plot between Working Hours Per Week and Income Level

E. Observations

None of the numerical attributes have missing values. The values are on different scales. Many machine learning models require the values to be on the same scale. We will use Standard Scaler from the sklearn library to scale the features.

F. Predictive Analysis Task

1) Definition - The task is defined as classification of income with two classes: >50k and <=50k.

2) Evaluation Method - The evaluation of models is done using the measure of accuracy on dataset.

3) Feature Selection - We used stepwise logistic regression to progressively eliminate features that weren't important. Following the elimination, only the important attributes were left and were divided into num and cat pipelines which were later combined to form a proper pipeline based on the principle of logical regression. For the choice of encoders used in converting pipelines we chose to use Label Encoder.

4) Baseline Performance - Our baseline model for this task is the accuracy when we predict every person in the data set to have an income of <=50k as that is the level which occurs more frequently.

With the above-mentioned prediction, our accuracy for the baseline is 0.89286.

G. Model Description

Using the scikit learn package in Python and Radiant package in R offers a wide variety of classifier models like Logistic, SVM, Neural Network, Random Forest, and various ensemble classifier techniques. It is a time-consuming process to determine which would be the best method to classify the data at hand. For our study, we decided to use the Logistic Regression Model.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression(or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name.

The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

H. Confusion Matrix

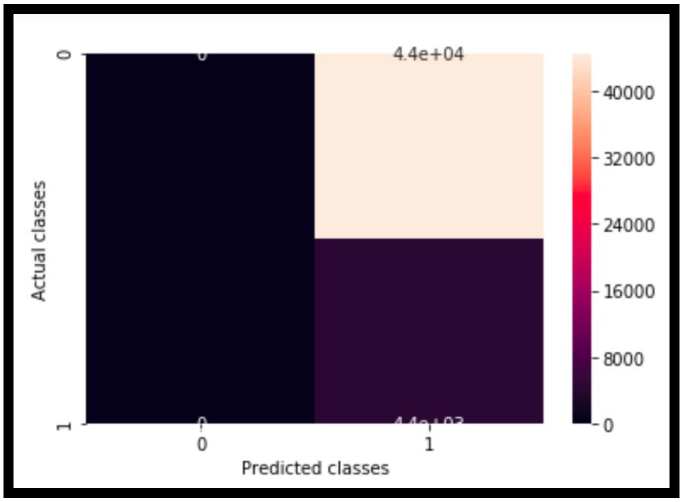


Fig. 9 Confusion Matrix between Predicted Classes and Actual Classes

The X-axis represents the Predicted classes and the Y-axis represents the Actual classes. How do we interpret the confusion matrix? 1.2e+04 times the model correctly predicted the class 0 when the actual class was 0. Similarly, conclusions can be drawn for the remaining cases.

IV. DISCUSSIONS

We used StratifiedKFold to divide our dataset into k folds. In each iteration, k-1 folds are used as the training set and the remaining fold is used as the validation. We use StratifiedKFold because it preserves the percentage of samples from each class.

If we use KFold, we might run the risk of introducing sampling bias i.e, the training set might contain a large number of samples where income is greater than 50K and test set contains more samples where income is less than 50K. Whereas StratifiedKFold will ensure that there are enough samples of each class in both the train and test dataset.

Finetuning – The model was finetuned using the GridSearchCV feature of sklearn.

The final accuracy score comes out to be 0.89286.

V. CONCLUSION

To conclude, we are encouraged by our results. Our classifiers are extrapolating patterns from the data, and this shows promise to be successfully can predict income based on Census information. An important takeaway is which classifiers worked best on classifying the Census data. While

SVM and KNN classifier worked best on some of the data that we randomly selected for train, it tended to perform much worse than the other son the test data. Making it unreliable, hence we would not recommend it for the task at hand. We received good results with artificial neural networks for 5 hidden layers and then incrementally better results for Gradient Boost classifier. As only 27% of our data's labels were >50K, boosting helped predict well for both the labels.

The future scope of this work involves achieving an overall 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.

VI. REFERENCES

[1] Chakrabarty, Navoneel & Biswas, Sanket. (2018). A Statistical Approach to Adult Census Income Level Prediction.

[2]Chandola, Tarani. "Social class differences in mortality using the new UK National Statistics Socio-Economic Classification." Social science & medicine 50, no. 5 (2000): 641-649.

[3]Alonso-Rodríguez, Agustín. "Logistic regression and world income distribution." International Advances in Economic Research 7, no. 2 (2001): 231-242.

[4]Stronks, Karien, H. Van De Mheen, Jill Van Den Bos, and Johan P. Mackenbach. "The interrelationship between income, health and employment status." International journal of epidemiology 26, no. 3 (1997): 592-600.

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

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

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

[8] 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.

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

[10] 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.

[11] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques"

[12] HaojunZhu:"PredictingEarning Potential using the Adult Dataset".