Adult Income Prediction Using various ML Algorithms

Sunil Thapa
Department of Computing
Lambton College
Toronto, Canada
C0846592@mylambton.ca

***Abstract*—This paper compares different Machine Learning Algorithms performances using an adult income dataset. Feature engineering, feature selection, and exploratory data analysis are performed to achieve this goal. Among the five Machine Learning Algorithms used, Random Forest Classifier performed best with 86.3% training accuracy and 86% test accuracy.**

Keywords*—Machine Learning, Random Forest Classifier, K-Neighbors Classifier, Logistic Regression, Naive Bayes Classifier, Support Vector Classifier, Confusion Matrix.

I. Introduction

The discipline of Artificial Intelligence has expanded significantly since Machine Learning Algorithms were introduced. These Algorithms have been used for various tasks, including research for classification and regression tasks. In addition to using them for research and discovery, data mining and machine learning fields have also explored specific obscure patterns and ideas that have enabled the prediction of difficult-to-forecast future events.

Here, the experiment was carried out to create a machine learning model from scratch to analyze data, extract/select its features, deal with missing values, and compare the performance of different algorithms. The dataset consists of 14 attributes viz age, workclass, fnlwgt (final weight), education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country.

II. Literature Review

Researchers have attempted to estimate income levels using machine learning models.

- For prediction tasks, including Logistic Stack on XGBOOST and SVM Stack on Logistic for scaling up the accuracy, Topiwalla [1] uses complicated techniques like XGBOOST, Random Forest, and stacking of models.

- Based on the 1994 Population Survey provided by the U.S. Census Bureau, Lazar [2] used Principal Component Analysis (PCA) and Support Vector Machine (SVM) methods to produce and assess income prediction data.

- Deepajothi et al. [3] attempted to reproduce Bayesian Networks, Decision Tree Induction, Lazy Classifier, and Rule Based Learning Techniques for the Adult Dataset and provided a comparative analysis of the predicted capabilities.

- In an attempt to simplify the complexity of several machine learning models used in classification tasks, Lemon et al. [4] sought to find the key features in the data.

- With the Ensemble Learning Algorithm and Gradient Boosting Classifier, Chakrabarty and Biswas [5] used Grid Search to tune the hyperparameter and gain an 88.16% validation accuracy.

III. Proposed Methodology

A. Dataset

The University of California Irvine (UCI) Machine Learning Repository was used to access the data for our analysis [6]. Ronny Kohavi and Barry Becker extracted the dataset from the 1994 census database. The data set contains information on 32,561 individual records and 14 attributes comprising six continuous variables and eight categorical features. The income level is a binary target label in the data set that indicates whether or not a person earns more than $50,000 per year based on the provided set of attributes.

B. Feature Engineering and Selection

In Fig. 1, a correlation matrix is displayed as a heat map with feature-to-feature correlation among the attributes, all of which are continuous variables. As seen in the heatmap, the attribute education and education-num are closely co-related, with 91% similarity. Hence, the attribute education-num is removed from the dataset to reduce the redundancy of feature values also known as performing an omission over the dataset.

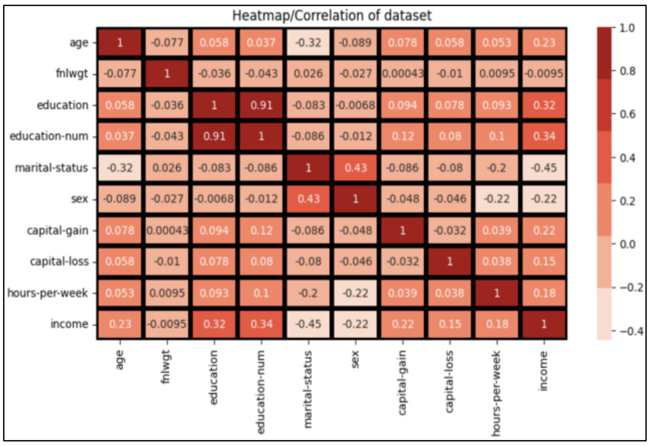

Figure 1: Heatmap

C. Exploratory Data Analysis

Figure 2, 3, 4, 5, 6 and 7 shows Kernel Density Estimate (KDE) and boxplots for continuous features to comprehend the measures of the attribute's central tendencies properly.

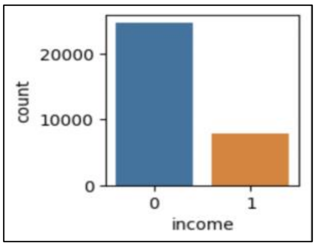

Figure 2: "Income" Count Plot

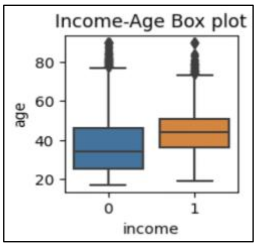

Figure 3: "Income" vs "Age"

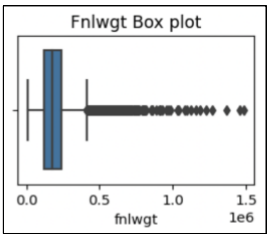

Figure 4: "Fnlwgt" Box Plot

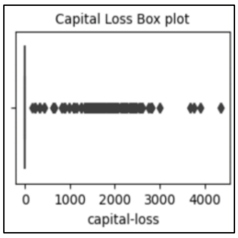

Figure 5: "Capital-loss" Box Plot

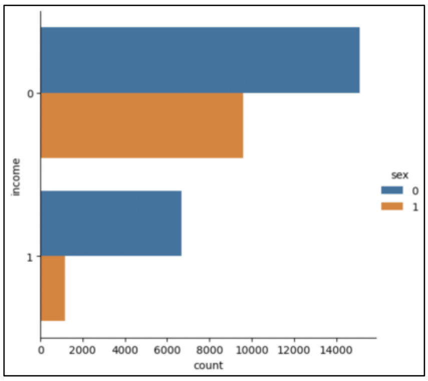

Figure 6: "Income Vs Sex"

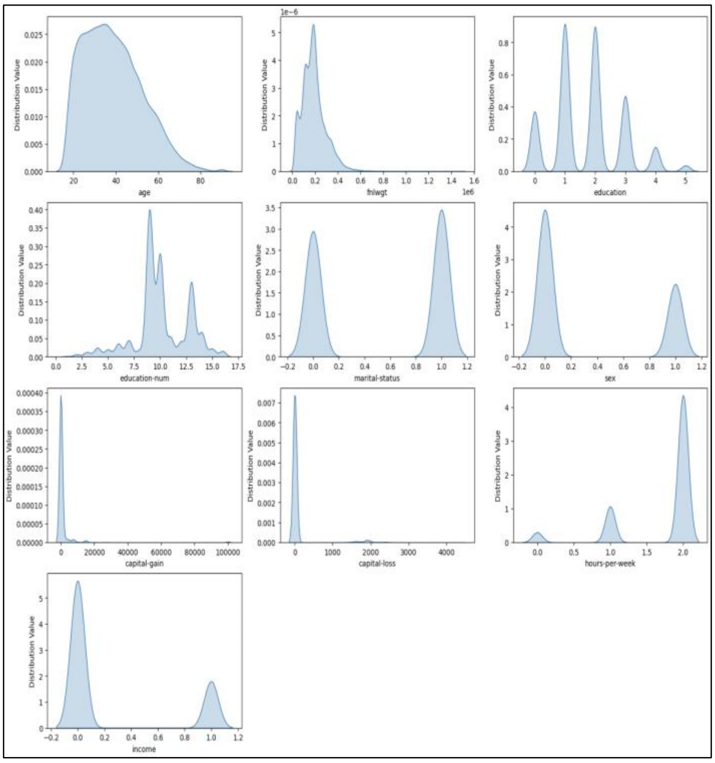

Figure 7: "Kernel Density Estimation (KDE)"

D. Data Pre-processing

The data must first be cleaned using preprocessing techniques before being processed using the Adult Dataset.

1) Handle Missing Values: Since the dataset comprises 32,561 records, a few missing values are represented as '**?**'. There were three attributes with such missing values, namely workclass, occupation and native-country. Since these are categorical values, the '?' were replaced with the most-occurred value.

2) Categorical Feature Encoding: There were few categorical attributes which could be grouped upon more and be encoded for better representation. Table 1 demonstrates the label encoding performed for such features.

Table 1

Feature	**Grouped Value**	**Label Encoded Value**
income	<=50k	0
	>50k	1
education	Preschool,1st-4th,5th-6th,7th-8th,9th,10th,11th,12th	0
	HS-grad	1
	Assoc-voc,Assoc-acdm,Prof-school,Some-college	2
	Bachelors	3
	Masters	4
	Doctorate	5
hours-per-week	< 20	0
	20 - 40	1
	> 40	2
marital-status	Married-civ-spouse,Married-AF-spouse	0
	Never-married, Divorced,Separated,Widowed, Married-spouse-absent	1
sex	Male	0
	Female	1

3) Train Test Split: The dataset is shuffled and split into training and test sets, with 70% of the data made available for training and the remaining 30% for testing.

E. Data Modelling

Several machine learning algorithms were used to perform the classification task for the Adult Income dataset. These include KNeighbors Classifier, Logistic Regression, Random Forest Classifier, Support Vector Classifier, and Naive Bayes. In order to determine the best hyper-parameter for these algorithms, an estimator search algorithm was implemented named Grid Search. Table 2 demonstrates the combination of hyper-parameter, the best parameter for each model, train and test accuracy.

Table 2: Hyper-Parameter Tuning Results

Classifier	**Hyper-parameter**	**Best parameter**	**Accuracy**	
			Train	**Test**

KNeighbors Classifier	'n_neighbors': range(1, 20, 2)	'n_neighbors': 15	84.5%	84%
Logistic Regression	'penalty': ['none', 'l2', 'l1', 'elasticnet'] 'C': np.logspace(-2, 0, 10) 'solver': ['newton-cg', 'lbfgs', 'sag', 'saga'] 'multi_class': ['multinomial']	C: 0.21544346900318834 multi_class: 'multinomial' penalty: 'l1' solver:'saga'	83.6%	84%
Random Forest Classifier	'max_depth': range(3, 20, 2) 'n_estimators': range(1, 15) 'max_features': [2, 3, 5, 7]	max_depth: 11 max_features: 5 n_estimators: 14	86.3%	86%
SVC	'C': [0.1, 1, 10, 100] 'gamma': [0.1, 0.01] 'kernel': ['rbf', 'sigmoid']	C: 1 gamma: 0.1 kernel: 'rbf'	84.6%	85%
Naive Bayes	default	default	81.9%	82%

Fig. 8 displays a graph summary of the Grid-Search Tuning of the KNeighbors Classifier, Logistic Regression and Random Forest Classifier based on their mean scores.

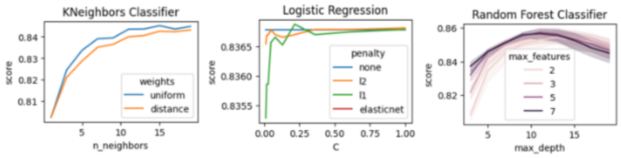

Figure 8: Graph Summary of Grid Search

IV. Results

In the dataset, there are 32,561 instances, out of which 22,792 were used for training, while the remaining 9,769 were set aside for testing. Random Forest Classifier (RFC) outperformed all others with 86% test accuracy among the tested Machine Learning Algorithms. The model performance is accessed using the following metrics:

- Precision is calculated by dividing the actual true prediction by the model's total number of predictions.

  Precision = TP/(TP+FP)

  RFC model resulted with precision of 0.86.

- The recall is determined in a classification problem with two classes by dividing the total number of true positives by the sum of true positives and false negatives.

  Recall = TP/(TP+FN)

  RFC model resulted with recall of 0.87.

- A weighted average of recall and precision is the F1 score.

  F1 score = 2(Recall * Precision) / (Recall + Precision)*

  RFC model resulted with f1-score of 0.86.

- Receiver Operating Characteristic Curve (ROC) is the derived from plotting True positive rate (TPR) against false positive rate (FPR) as shown in figure 9.

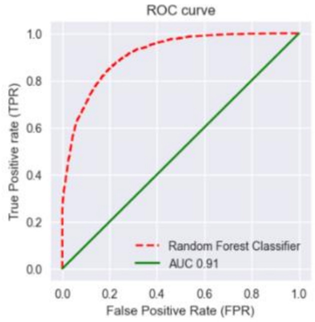

Figure 9: ROC curve

- The classification model's success in correctly predicting examples from different classes is summarised in a table called the confusion matrix. The confusion matrix generated by Random Forest Classifier is shown in figure 10.

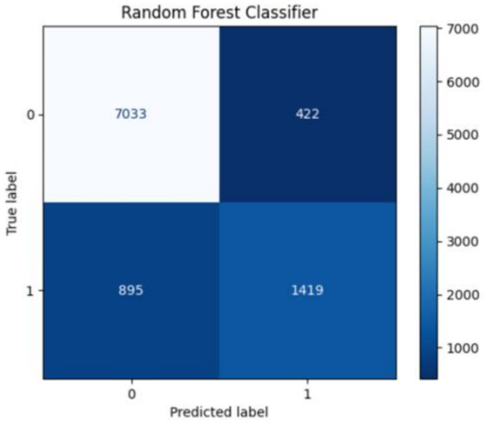

Figure 10: Confusion Matrix

V. Conclusion

As per the study, Random Forest Classifier demonstrated higher accuracy compared to other mentioned algorithms. As per the confusion matrix, 895 were classified as False Positive, while 422 were classified as False Negative. Hence, 13-14% of the test data are errors produced by the model. In order to improve the accuracy of the model, the extended version of the current dataset having 48,842 records can be used, as Chakrabarty and Biswas [5] obtained 88.16% validation accuracy on an 80-20 train-test split. Therefore, future work on this project will focus on implementing hybrid Artificial Intelligence techniques - a combination of Machine Learning and Deep Learning (Neural Networks) - to produce better results overall while maintaining accuracy.

References

[1] Topiwalla, M. (n.d). *Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting.* DataScience. https://datascience52.files.wordpress.com/2017/02/machine-learning-

on-uci-adult-data-set-using-various-classifier-algorithms-and-scaling-up-the-accuracy-using-extreme-gradient-boosting.pdf.

[2] Lazar, A. (2004). Income prediction via support Vector Machine. *2004 International Conference on Machine Learning and Applications, 2004. Proceedings.,* 143-149. https://doi.org/10.1109/icmla.2004.1383506.

[3] Deepajothi, S., Selvarajan, S. (2012). A Comparative Study of Classification Techniques On Adult Data Set. *International Journal of Engineering Research & Technology (IJERT)*, 1(8).

[4] Lemon, C., Zelazo, C., Mulakaluri, K. (n.d). *Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques.* CSEWEB.UCSD.EDU. https://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf

[5] Chakrabarty, N., Biswas, S. (2018). A Statistical Approach to Adult Census Income Level Prediction. *arXiv*. https://doi.org/10.48550/arXiv.1810.10076

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