Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230111

Performances evaluation of machine learning models on income forecasting

Ziyang Wan

School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, 710126, China

21009201316@stu.xidian.edu.cn

Abstract. Job seekers, especially those who are looking for their first job, often lack sufficient experience and guidance, which makes it difficult for them to obtain satisfactory salaries. Therefore, salary prediction is very important. For individuals, income ranges can be estimated; for companies, the use of such estimates can guide salary adjustments for employees and prevent the loss of talented personnel, increase company revenue, and reduce operating costs; for governments or countries, these estimates can provide a macro-level assessment of overall income for a large area, such as predicting GDP per capita in a city, making it easier to make economic adjustments and grasp macro development trends. This article uses three models: decision trees, random forests, and neural networks, to train relevant datasets. The dataset is Adult Income Dataset from Kaggle. A total of 32,561 adults are included, including 15 items of data including age, education level, occupation, marital status, working hours per week, and others. The training and test sets were divided into a 7:3 ratio, and the predictive result of each model was evaluated through following figures: accuracy, recall rate, and F1 score. The final conclusion was that the random forest model had the best performance. There is an inseparable relationship between residents' income and the development and happiness of individuals and social stability.

Keywords: machine learning, income prediction, neural network.

1. Introduction

With economy developing, residents' income level has become a focus of social attention [1]. Income prediction mainly involves mining basic information about residents to predict their income situation [2, 3]. In this study, different models were used for machine learning to obtain a more accurate predictive function for adult income, and the learning performance of each model was also compared. At present, relevant income forecasting occupations are relatively rare, and artificial forecasting may be affected by subjective factors. Using machine learning for prediction has many advantages [4, 5]. Firstly, it has faster predictions, and a lot of time can be saved for data reading and calculation. Secondly, the algorithms have very low economic costs. Compared with manual positions, the use of machine calculations will bring little economic expenditure to the company. Thirdly, it saves manpower and enhances the company's production capacity. Fourthly it processes strong learning ability. Manual prediction may take a long time to learn to make a more reliable judgment, but using a machine can quickly learn and build a model. Fifthly, it has strong adaptability, as long as the

DOI: 10.54254/2755-2721/27/20230111

corresponding data is obtained, models, can be constructed, regardless of the reality of time or space. Finally, the prediction result is objective and will not be affected by subjective factors.

2. Method

2.1. Dataset

A total of 15 relevant data of 32561 adults were included. It is divided into training and testing set with a ratio of 7:3. Features and their corresponding importance are demonstrated in Figure 1 and Figure 2 respectively. Dataset is downloaded from Kaggle [6].

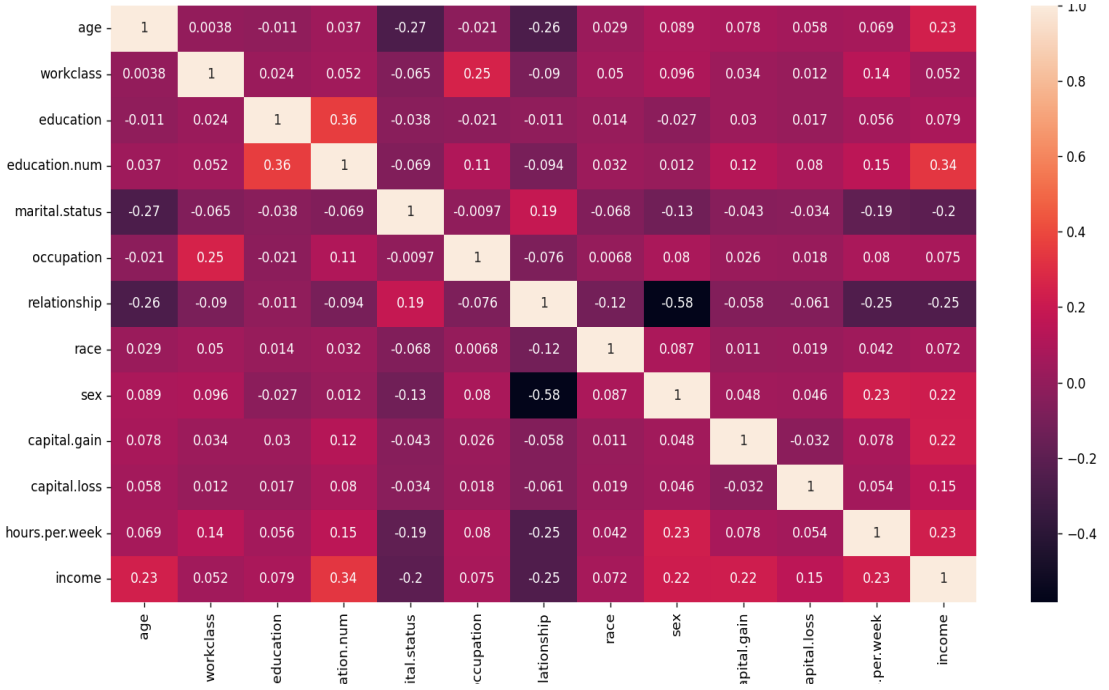

Figure 1. Heat map of feature correlations.

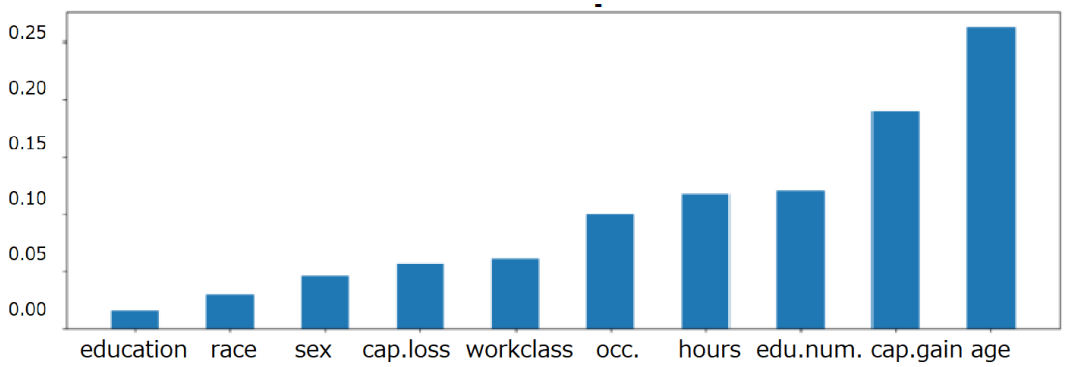

Figure 2. Feature importance.

2.2. Data preprocessing

Data preprocessing consists of three steps. In the first step, it is observed that the data types include integers and characters. To facilitate machine operations, LabelEncoder is used to convert all character data into numeric data. The second step is to conduct a correlation analysis on all the data and delete the data with weak correlation, so as not to affect the accuracy of the model. Table 1 shows the 13 remaining variables after removing fnlwgt (final weight) and native.country.

DOI: 10.54254/2755-2721/27/20230111

Table 1. Selected variables with strong correlation.

variable name	meaning	Data type (raw data)
age	age	Integer(17~90)
workclass	working class	character
education	Highest education	character
Education.num	school year	Integer(1~16)
Marital.status	Marriage	character
occupation	Occupation	character
relationship	family relationship	character
race	race	character
sex	sex	Integer(0,1)
Capital.gain	capital gains	Integer(0~100000)
Capital.loss	capital loss	Integer(0~4356)
Hours.per.week	Working hours per week	Integer(1~99)
income	Income	character（<=50K，>50k）

The third step is normalization, which aims to make the influence of each data point more similar. The main purpose of this step is to improve the accuracy and stability of the algorithm: in machine learning, many algorithms depend on the scale and range of the data. Normalizing the data can put features with different units on the same scale, avoiding the influence of feature sizes on the algorithm, and thereby improving the accuracy and stability of the algorithm. Secondly, it can accelerate the convergence speed of the model where data normalization can make the optimization algorithm converge faster because it can eliminate the numerical differences between different features, thereby reducing the magnitude of gradient updates and improving the efficiency of the optimization algorithm.

2.3. Neural network

Neural networks are a computational model that simulates the biological neural system, and they are a type of machine learning algorithm. Initially designed to simulate the neurons, synapses, and signal transmission in the brain, neural networks have undergone multiple evolutions and improvements and have now become one of the most widely used algorithms in the fields of not only machine learning but also deep learning.

The study of neural networks began in the 1940s, with the earliest neuron models being proposed. However, it was not until the 1980s, with the improvement of computer performance, that research and application of neural networks gradually emerged. In recent years, with the rise of deep learning, neural networks have been used in many fields, for example, optical character recognition, speech recognition and natural language processing.

The neural networks have three main advantages [7, 8]. 1) on-linear relationships can be learned and handled. Compared with traditional machine learning algorithms, neural networks can automatically learn more complex nonlinear relationships from data, and can handle large amounts of complex data, such as images, voice, text, etc. 2) It has good adaptability and generalization ability. The neural network can automatically adjust the weights and parameters of the model, and can adapt to different data sets and environments. At the same time, it has good generalization ability and can deal with unknown data. 3) Can handle large-scale data and high-dimensional data. As the scale of data increases, neural networks can learn more accurate models from large-scale data, and can also process high-dimensional data, which is suitable for complex pattern recognition problems.

The principle of neural networks is mainly to simulate the information transmission process between neurons, which are connected by multiple neurons according to a certain topological structure. Each neuron receives some input signals and processes them through an activation function to output a result, while also passing this result to the next layer of neurons. The learning process of a neural

DOI: 10.54254/2755-2721/27/20230111

network can be implemented through the backpropagation algorithm, which uses gradient descent to update the weights and parameters of the neural network.

In the training of the three models, the settings of the neural network including three hidden layers, containing 40, 20, and 1 neuron respectively.

2.4. Decision tree

Decision Tree is a machine learning algorithm based on tree structure for decision analysis. It analyzes the features of data and builds a tree structure to classify or predict new data. In a decision tree, each node express a characteristic, branches express possible values of the characteristic, and leaf nodes express classification results. When a new data sample is classified by the decision tree, it starts from the root node and goes down the tree structure until it reaches a leaf node, and then determines the category of the new data based on the classification result of the leaf node. The model is widely used in many practical applications, such as medical diagnosis, financial risk assessment, industrial quality control, market forecasting and other fields [9].

The algorithm has three advantages. 1) Easy to understand and explain: The structure of the decision tree can be visualized, which is easy for humans to understand and interpret, and can intuitively show the impact of different features on the results. 2) Can process high-dimensional data and nonlinear relationships: the decision tree can effectively process high-dimensional data and nonlinear relationships. Decision tree can gradually reduce the feature space through the branch structure of the tree, thereby reducing the impact of the number of features on decision-making. 3) Can handle mixed data types: the decision tree algorithm can handle mixed data types containing numeric, categorical, and Boolean features without feature conversion. 4) Can handle missing values: When constructing a decision tree, different strategies can be used to handle missing values, such as pruning, substitution and other methods.

2.5. Random forest

Random Forest is a common machine learning algorithm that belongs to the ensemble learning. It is a model composed of multiple decision trees, where each tree is a classifier. When faced with classification problems, the export of the random forest is the pattern of the export of all decision trees, while in a regression problem, the output is the average of the outputs of all decision trees. Each decision tree in a random forest is obtained by bootstrapping the data set. In bootstrap sampling, a new dataset is formed by randomly sampling a certain number of samples with replacement from the original dataset. The size of the new dataset is the same as the original dataset, but there may be repeated samples. Each decision tree in a random forest is trained on a different bootstrap sample dataset. During the training of each decision tree, a random subset of features is also selected for splitting. This method is known as random feature selection [10].

The advantage of random forest is that it can handle high-dimensional data well, and has some robustness to missing data and noise. In addition, random forest is not prone to overfitting and can effectively handle non-linear relationships. Therefore, random forest is a commonly used machine learning algorithm and has a wide range of applications in many practical problems.

3. Result

Three models were trained on the dataset and tested to evaluate their learning and fitting performance. The results in Table 2 show that the recall rate for the decision tree and neural network models is 0.72, which is lower than that of the random forest model (0.77). The decision tree and neural network models have similar precision and positive F1-score, but the random forest model has a clear advantage. In terms of negative F1-score, all three models are very close. Therefore, the analysis suggests that the random forest model has the best fitting performance and is suitable for predicting adult income.

DOI: 10.54254/2755-2721/27/20230111

Table 2. Performance indicators of three models.

	Recall	Accuracy	+F1-score	-F1-score
Decision Tree	0.72	0.796	0.59	0.88
Random Forest	0.77	0.815	0.63	0.89
Neural Networks	0.72	0.798	0.57	0.88

4. Discussion

In the model exploration, the neural network did not show outstanding performance, which may be due to the influence of the dataset form. Neural networks are suitable for three-dimensional model processing, such as image recognition. However, this dataset is a two-dimensional dataset, and the correlations between data are weak, so the performance of the neural network did not fully play out. The random forest showed the best performance. This is possible because random forest improves the predictive accuracy of a single decision tree by combining the predictions of multiple decision trees and avoiding overfitting by randomly selecting training samples and features. Each decision tree is trained on a different bootstrap sample and only uses a subset of features for splitting. In this way, each decision tree has a certain degree of independence and can capture different features of the data. Ultimately, the prediction result of random forest is based on the voting or average of the output results of all decision trees. Therefore, random forest is an algorithm built on decision trees that inherits some of the advantages of decision trees and improves predictive accuracy by integrating the predictions of multiple decision trees, while having better robustness and interpretability.

According to the model's predictions on feature importance, it was found that the most influential factor on income is the number of years of education. Therefore, it can be concluded that the most effective way to increase one's income is to increase the number of years of education and improve one's education and skills. This also matches with people's perception and common knowledge.

As age increases, the proportion of high-income individuals also increases. However, there is no direct correlation between the independent and dependent variables. This can be comprehended by the truth that older individuals are more possible to have longer education experiences and more work experience and life experiences, which may contribute to the increase in the proportion of high-income individuals.

5. Conclusion

In this works three algorithms are conducted in predicting the income of. In the training of models for predicting adult income, three models are decision tree, random forest, and neural network. The data set was split into a 7:3 ratio of training set and testing set for training. The random forest model achieves the best performances. Based on the research on this dataset, it was found that the random forest model had the best fitting performance. Therefore, when predicting income, it is recommended to prioritize trying the random forest model to obtain more accurate prediction results. In addition, it could be found that if people want to increase their income, they can enhance their education through training and learning to play a more important role in their work. They can also increase their working hours appropriately to achieve income growth.

References

[1] Kibekbaev, A., & Duman, E. (2016). Benchmarking regression algorithms for income prediction modeling. Information Systems, 61, 40-52.

[2] Leppäniemi, H., Ibrahim, E., Abbass, M. M., Borghi, E., Flores-Urrutia, M. C., et, al. (2023). Nutrition Profile for Countries of the Eastern Mediterranean Region with Different Income Levels: An Analytical Review. Children, 10(2), 236.

[3] Allen, L., Williams, J., Townsend, N., Mikkelsen, B., et, al. (2017). Socioeconomic status and non-communicable disease behavioural risk factors in low-income and lower-middle-income countries: a systematic review. The Lancet Global Health, 5(3), e277-e289.

DOI: 10.54254/2755-2721/27/20230111

[4] Saavedra, M., & Twinam, T. (2020). A machine learning approach to improving occupational income scores. Explorations in Economic History, 75, 101304.
[5] Dutt, P., & Tsetlin, I. (2021). Income distribution and economic development: Insights from machine learning. Economics & Politics, 33(1), 1-36.
[6] Aditi, M. (2021) Adult Income Dataset, URL: https://www.kaggle.com/code/aditimulye/adult-income-dataset-from-scratch/data
[7] Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., et, al. (2019). Investigate neural networks!. J. Mach. Learn. Res., 20(93), 1-8.
[8] Bishop, C. M. (1994). Neural networks and their applications. Review of scientific instruments, 65(6), 1803-1832.
[9] Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39, 261-283.
[10] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.