Income Level Prediction

Constantin Amurov$^{[BPBLNO]}$

Eötvös Loránd University, Budapest 1053, Hungary
https://www.elte.hu/en/

Abstract. In this report we will use a number of different supervised algorithms to precisely predict individual's income using Adult Data Set collected from the UCI Machine Learning Repository. Then we will choose the best algorithm from results. After this we will the K-Means clustering algorithms to correctly cluster the data. Our goal is to build a model that accurately predicts whether an individual makes more than 50000$.

Keywords: Feature Engineering · Model Building · Logistic Regression Classifier · Random Forest · Confusion Matrix · K-Means Clustering. · Elbow Method · Frequent Pattern Mining · Apriori Algorithm

1 Introduction

The data set that we use is the US Adult Census data set provided by UCI Machine Learning Repository.It consists of 32560 entries that had been extracted from the 1994 US Census database.

1.1 Exploratory Data Analysis

The data set consists of 32560 entries. Each entry contains a mix of categorical and numeric data type that provides information about the individual.

- **Age** - Integer greater than 0. represents the age of the individual
- **Work Class** - Categorical Data. Represents the employment status of the individial.
  - Private, Self emp not inc, Self emp inc, Federal gov, Local gov, State gov, Without pay, Never worked.
- **fnlwgt** - Represents the number of the the census believes the entry represents
- **Education** - Categorical data. Represents the highest level of education achieved by an individual.
  - Bachelors, Some college, 11th, HS grad, Prof school, Assoc acdm, Assoc voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **Education Num** - Represents in numerical form the highest level of education achieved by individual

- **Marital Status** - Represents the marital status of the individual.
  - Married-civ-spouse - Civilian spounse
  - Married-AF-spouse - Spouse in the Armed Forces.
- **Occupation** -Represents the occupation of the individual
  - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **Relationship** -Represents the individual's relation to others.
  - Wife, Ownchild, Husband, Notinfamily, Otherrelative, Unmarried.
- **Race** -Represents the race of the individual
  - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **Gender**
  - Male
  - Female
- **Capital Gain** -Continuous data. Represents the capital gain of the individual
- **Capital Loss** -Continuous data. Represents the capital loss of the individual
- **Hours Per Week** -Continuous data. Represents the hours an individual has reported to work per week
- **Native Country** -Country of origin of the individual
  - 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, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, TrinadadTobago, Peru, Hong, Holand-Netherlands.
- **Income** - Label that indicates whether the individual makes more than 50000$ per year or not

Correlation

Heat Map. Using the Heat Map Technique we can conclude that there is no any important correlation between the features of the individuals (see Fig. 1). The greatest correlation is between the **education number** of the individual and its **income**.

Race and Income. Using the plot we can see that the race is linked to the income. Our plot counts that more than 20000 individuals of white race have an income greater than 50000 $, whether only more than 2500 individuals of black race have an income greater than 50000 $ (see Fig. 2).

Distribution

Box Plot. Using the box plot method we can see that the majority have a capital gain between 0 to 40000 $ per year. The average numbers of hours per week are 40 and education number is between 9 to 12(see Fig. 4).

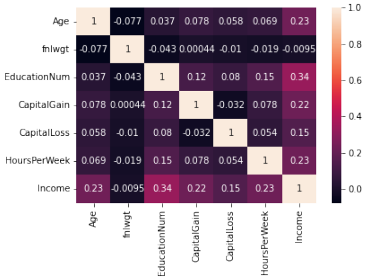


Fig. 1. The correlation of features of individuals based on Heat Map Technique

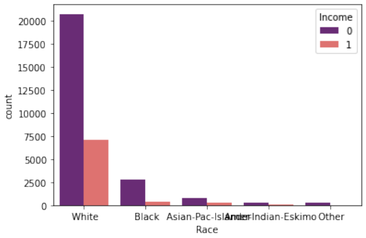


Fig. 2. The distribution of income based on the race

Gender and Income. Using the plot we can see that the gender is linked to the education. Our plot counts that more than 6500 individuals of male gender have higher education level, whether only more than 3900 individuals of female gender have higher education level (see Fig. 3).

2 Prediction task

For our prediction we will opt for Supervised Learning and exactly 3 types of Classification Algorithms.

- Logistic Regression Classifier
- Random Forest
- Decision Tree Classifier

Before everything else, we will use Feature Engineering principles to preprocess our data set. After this we will build our model.

2.1 Feature Engineering

Because many of education categories are redundant for our prediction, we reduced 16 education categories to only 6. The same principle we applied for the

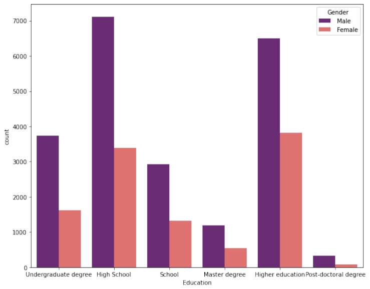


Fig. 3. The distribution of education based on the gender

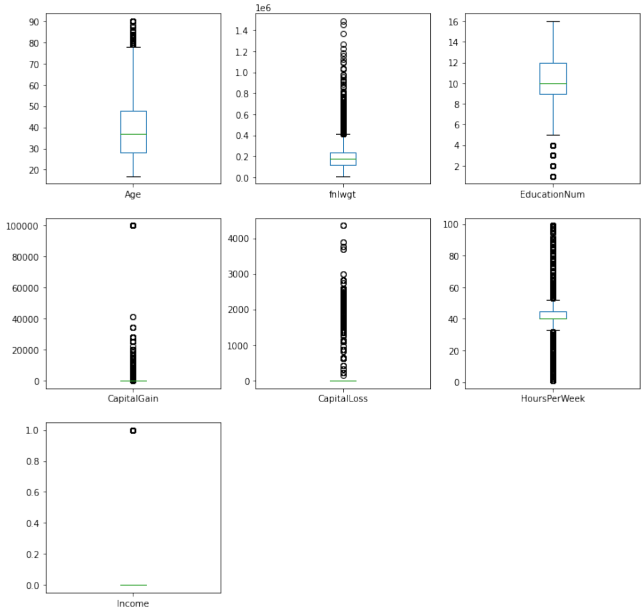


Fig. 4. The distribution of features of the individuals based on Box Plot method

marital status of the individuals, we reduced 6 marital status categories to only 2. As our output data consists of only two categories (>50K,≤50K), we transformed the data in a more simple representation (0,1)

2.2 Feature Scaling

For a better accuracy we used the standardization principle on features using StandardScaler method. After that, we applied encoding principles using LabelEncoder method. Our data set is ready for applying prediction algorithms.

2.3 Logistic Regression Classifier

Using the Confusion Matrix table, the algorithm ran on the training data has shown an accuracy of **0.841**, when the accuracy of test data has shown very similar results **0.843**.(see Fig. 5).

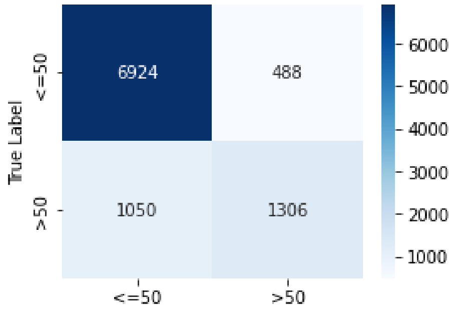


Fig. 5. Confusion Matrix - Logistic Regression Classifier

2.4 Random Forest Classifier

Using the Confusion Matrix table, the algorithm ran on the training data has shown an impressive accuracy of **1.000**, when the accuracy of test data has shown a result of **0.860**.(see Fig. 6).

2.5 Decision Tree Classifier

Using the Confusion Matrix table, the algorithm ran on the training data has shown an impressive accuracy of **1.000**, when the accuracy of test data has shown a result of **0.812** (see Fig. 7).

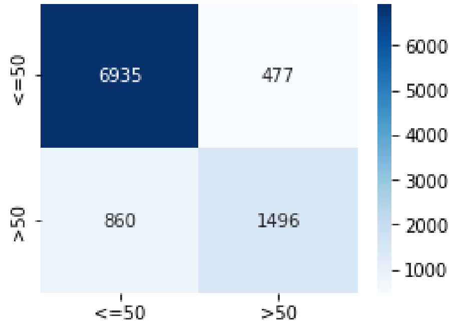


Fig. 6. Confusion Matrix - Random Forest Classifier

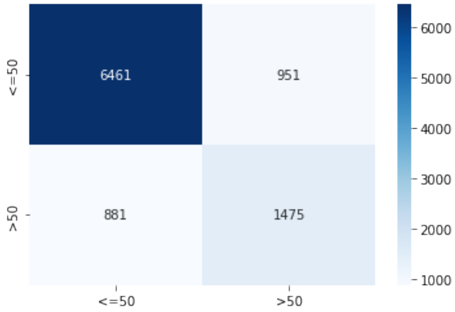
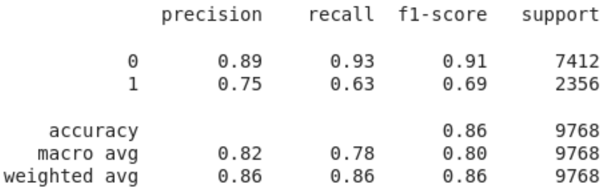


Fig. 7. Confusion Matrix - Decision Tree Classifier

Confusion Matrix We will use the model created by Logistic Regression Classifier for creating our confusion matrix (see Fig. 8). because it got good results. We got a precision of **0.88%** and a recall of **0.93%**

$$Precision = tp/tp + fp \tag{1}$$

$$Recall = tp/tp + fn \tag{2}$$



```
              precision    recall  f1-score   support

           0       0.89      0.93      0.91      7412
           1       0.75      0.63      0.69      2356

    accuracy                           0.86      9768
   macro avg       0.82      0.78      0.80      9768
weighted avg       0.86      0.86      0.86      9768
```

Fig. 8. Confusion Matrix

3 Clustering task

For clustering we will use K-Means clustering as it is a simple yet powerful algorithm.

3.1 Elbow Method

For figuring out what is the most optimal number of clusters, we used the elbow method. We iterated from 0 to 15 clusters on the data set to figure out what is the number of clusters that has the best number of sum of squared distances of samples to the nearest cluster centre. Our calculations shows that the best number of clusters is k=2 (see Fig. 9).

3.2 Plotting

The plot clearly shows the two clusters correctly predicted using the actual and predicted values on the testing data set (see Fig. 10).

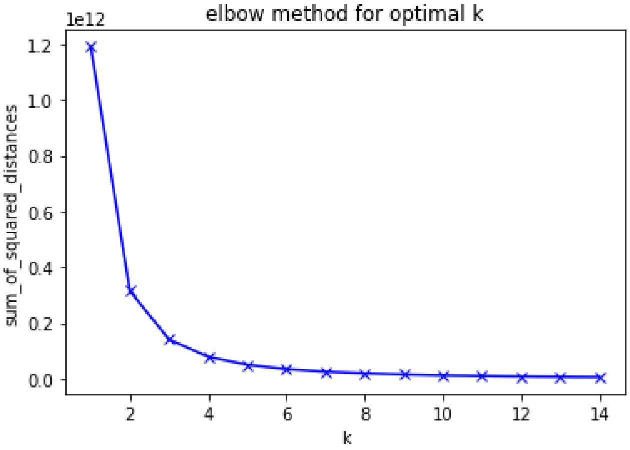


Fig. 9. Elbow Method

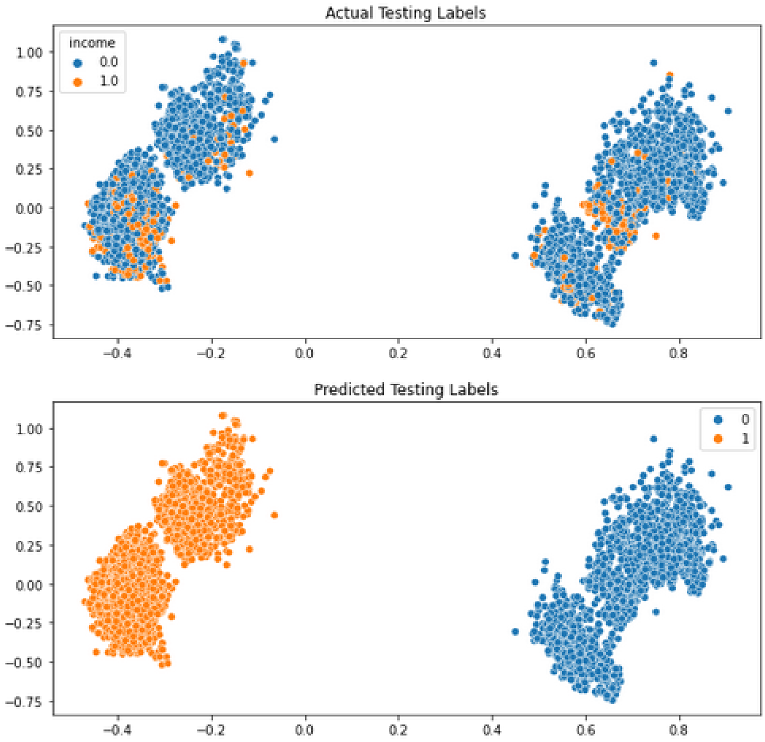


Fig. 10. K-Means Clusters Plotting

4 Frequent Pattern Mining

It is very well known that, finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. After running the Apriori Algorithm on our data set it has shown some interesting patterns like:

- (No Gain, No Loss, Full time, Private)
- (No Gain, No Loss, United-States,Private)
- (No Gain, No Loss, Private, White)
- ,(No Gain , Full time, United-States, White)
- (No Gain, United-States, Private, White)
- (No Loss, Full time, United-States, White)
- (No Loss, United-States, Private, White)

From this data we can make a clear association that No Loss label that stands for capital loss of 0 , is linked to Full time work of an individual of race White. Now we can make a prediction that individuals of race White are more likely to not have a capital loss. Moreover, this data can help us in clustering. For example individual of race White are more likely to have a full time job and have a work class of Private, we could use this data to enhance clustering precision.

5 Conclusion

In conclusion, we successfully created several supervised Machine Learning models that predicted the whether an individual has an income greater then 50000 $ per year or not. Moreover, we successfully calculated the optimal number of cluster for the data set and used the K-Means clustering algorithm for creating the clusters.

In addition to this, we concluded that Random Forest Classifier had the greatest accuracy comparing to other classifier used in the report , the algorithm has shown an impressive accuracy of 1.000, when the accuracy of test data has shown a result of 0.860.