Yearly income range prediction based on census data

Donia Gharbi

Eötvös Loránd University, Faculty of informatics, Budapest, Hungary
qnghnc@inf.elte.hu

Abstract. Annual payments and salaries depend on a lot of factors like sex , age , nationality, education ... also the time era makes a difference. In this project we will use the "Census Income" dataset to predict whether someone's income exceeds fifty-thousands dollars a year or not based on census data.

Keywords: Introduction to data science · Census income.

1 Introduction

An annual salary is the amount of money an employer pays their employee over the course of a year in exchange for the work they perform. The salary they receive is based on a lot of factors like sex , education , nationality ... In this paper we will work with a specific data set that has multiple features about a diverse list of employees and try to predict if their income exceeds fifty-thousands dollars a year or not.

First off we will go through the Exploratory Data Analysis where we discovered our dataset then Data pre-processing where we will make sure that our dataset will give the best results through some steps then Classification and Clustering Where we will go through two model examples for each of them and compare the results and lastly we will gave a brief Frequent pattern mining description.

2 Exploratory Data Analysis

2.1 About data set

This data set called "Census Income" was extracted by Barry Becker from the 1994 Census database, which makes the conclusions we will get here not accurate to the current generation due to the time gap. The target variable is the income which is a categorical variable defined by either '$<= 50K$' or '$> 50K$'.

The dataset has 32561 credit record and fifteen features , fourteen of them are predictors and one is the outcome variable which will be defined in the following table :

Table 1. The description of the attributes of the dataset

Name	Type	Outcome
income	Categorical	Outcome
workclass	Categorical	Predictor
education	Categorical	Predictor
marital-status duration	Categorical	Predictor
occupation	Categorical	Predictor
relationship	Categorical	Predictor
race	Categorical	Predictor
native-country	Categorical	Predictor
sex	Categorical	Predictor
fnlwgt	continuous	Predictor
capital-gain	continuous	Predictor
age	continuous	Predictor
capital-loss	continuous	Predictor
hours-per-week	continuous	Predictor
education-num	continuous	Predictor

3 Data pre-processing

3.1 Clean data

We checked if there were any null values and it turns out we have none.Then we replaced every '?' entry with Null using the function and checked the null values again and it turns out that there are 1836 ones in the workplace, 1843 in occupation and 583 in native-countries, to fix it we used the fillna() method that replaces the NULL values with a specified value.

3.2 Encoding

Encoding or continuization is the transformation of categorical variables to binary or numerical counterparts.We have used label encoding as a modeling method.

First off, before encoding our data , since our target variable is income and it's a categorical value , we have to switch it to a numerical one. In our case we replaced the value '$<= 50K$' with 0 and the value '$> 50K$' with 1.

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

In our case we separated the categorical and numerical values as they were specified in the description of the attributes of the dataset table (Table 1), the income now counts as a numerical value now since it has been switched to one previously. and then we encode the categorical values.

3.3 Standardization

Standardization is a scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

StandardScaler StandardScaler performs the task of Standardization. Usually a dataset contains variables that are different in scale. Four example ,in our case an our dataset contains age column with values on scale 20-70 and capital-gain column with values on scale 10000-80000. As these two columns are different in scale, they are Standardized to have common scale while building machine learning model.

3.4 Feature selection

Feature Selection is a method for reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of selecting relevant features for your model based on the problem you are trying to solve.In our case we chose Chi-Square to test the Significance of the attributes we ended up taking out the fnlwgt attribute.

The Chi-Square The Chi-Square Statistic is a number that describes the relationship between the theoretically assumed data and the actual data. It is usually considered as a number or statistic value that verifies the theoretical dataset with the actual dataset and gives the result in the form of a number.

4 Classification

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

4.1 Splitting dataset

We have to split our dataset into two datasets one for training and the other for testing.

the training dataset is the sample of data used to fit the model and the testing dataset is The sample of data used to provide an unbiased evaluation of a model fit on the training dataset.

In our case we will use our pre-processed data to predict if the income their income exceeds fifty-thousands dollars a year or not.

We start off by dropping our target class 'income' from the dataset and store it in another value.

Then when we split our data using the train_test_split method, the methods returns train and test datasets in the ratio of '75 : 25'.

Often, we want to preserve the dataset proportions for better prediction of results ,Since in our dataset income has a huge gap in between the entries for both values as in shows in the figure below: 75,92% for '$<= 50K$' and 24,082% for '$> 50K$' we have to make sure that we preserve the proportion of target as in original dataset, in the train and test datasets as well, so we used stratify parameter to ensure the income data proportion is preserved.

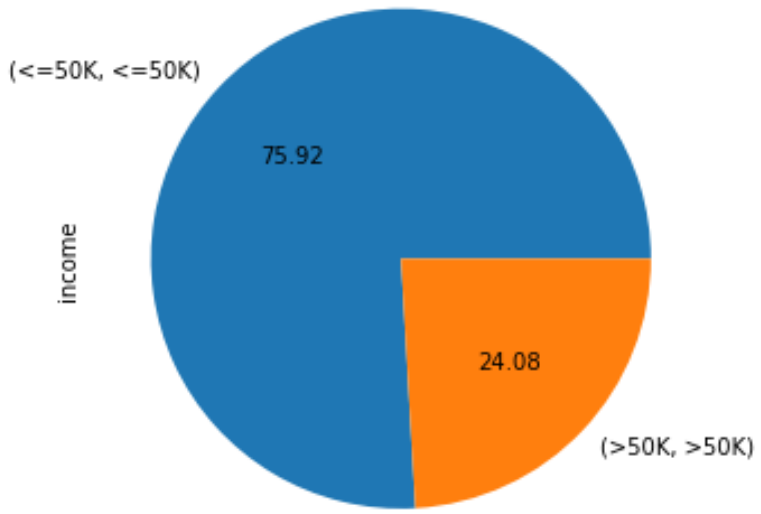

Fig. 1. Income percentage

4.2 Classification Models

For this part we have chosen two models Naive Bayes and Logistic Regression

a. Tuning Hyperparameters for Classification
Machine learning algorithms have hyperparameters that allow you to tailor the behavior of the algorithm to your specific dataset.

Hyperparameters are different from parameters, which are the internal coefficients or weights for a model found by the learning algorithm. Unlike parameters, hyperparameters are specified by the practitioner when configuring the model. To make sure that we set our algorithms with the best options possible for hyperparameters we used GridseachCV to specify their values.

GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

b. Naive Bayes
Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

c. Logistic Regression
Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

4.3 Training

a.Naive Bayes
Using Naive Bayes before tuning it's parameters we got the the Accuracy score 80.83 % and after tuning its' parameters we got the following score 81.76 % .

b.Logistic Regression
Using Logistic Regression before tuning it's parameters we got the the Accuracy score 80.83 % and after tuning its' parameters we got the following score 82.42 %.

4.4 Testing

Based on the mentioned numbers in the Training part we ended up choosing Logistic Regression as a model in the testing phase since it has a higher Accuracy score and we ended up getting the following score using the test dataset 81.92 % and its' performance is portraied through a confusion matrix.

Confusion matrix it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with four different combinations of predicted and actual values. the four different combinations are True Positive , True Negative , False Positive and False Negative.

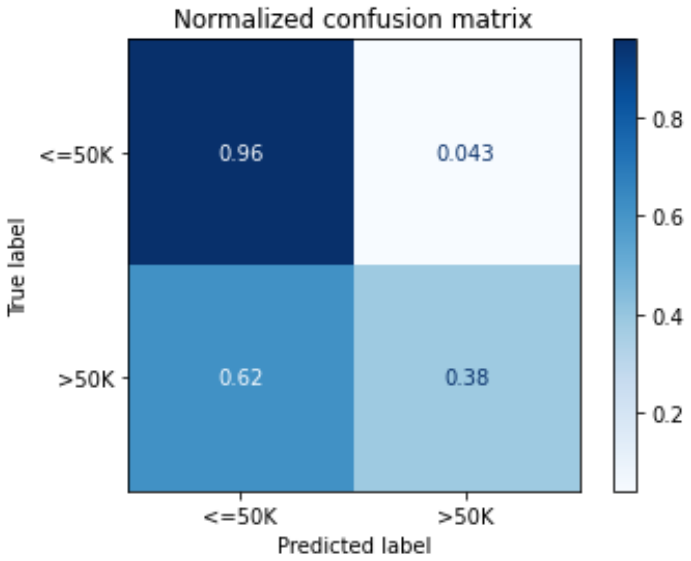

Fig. 2. Nomalized confusion matrix

5 Clustering

Clustering is the grouping of objects such that objects in the same cluster are more similar to each other than they are to objects in another cluster. The classification into clusters is done using criteria such as smallest distances, density of data points, graphs, or various statistical distributions.

5.1 Testing

a. KMeans

K-means is a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

In order to get the optimal result from the K-means algorithm we used the Elbow Method to get the optimal number of clusters which is the Hyperparameter for K-means.

b. Agglomerative Clustering

Agglomerative Clustering The Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting). It's a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

In order to get the optimal result from the Agglomerative Clustering algorithm we used the hierarchy dendrogram to get the optimal number of clusters which is one of the Hyperparameters for it.

5.2 Result

a. K-Means
To measure how efficient K-Means is in our case as a clustering method , we used Silhouette score calculation and the score is 30.1 %

Silhouette Coefficient Silhouette Coefficient or silhouette score is a metric used to calculate a clustering technique is. Its value ranges between -1 and 1.
In the figure bellow a visualisation of the three clusters we got could be seen , there is no overlapping in between the clusters. We got the following number of records in each cluster:

- 2 : 15246
- 1 : 15229
- 0 : 2086

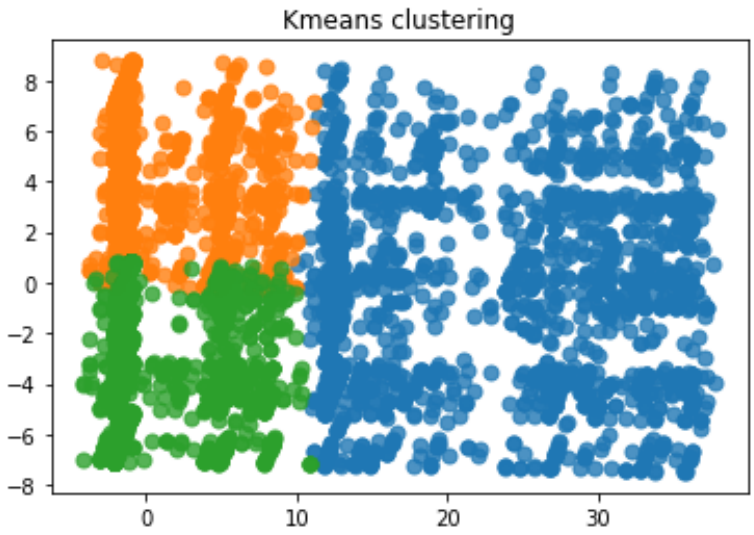

Fig. 3. K-Means clusters

b. Agglomerative Clustering

To measure how efficient Agglomerative Clustering is in our case as a clustering method , we used Silhouette score calculation and the score is 22.57 %
In the figure bellow a visualisation of the three clusters we got could be seen , there is some overlapping in between the clusters. We got the following number of records in each cluster:

- 2 : 10803
- 1 : 965
- 0 : 20793

Conclusion Based on both results from the Agglomerative Clustering and K-Means Clustering we can see that K-Means works better on our dataset based on the Silhouette score and the overlapping seen in the visualization of both clustering results.

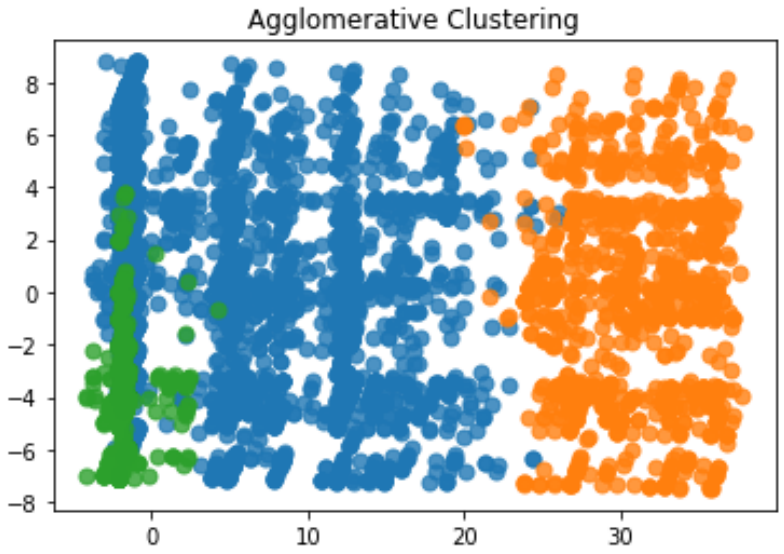

Fig. 4. Agglomerative Clustering clusters

6 Frequent pattern mining

6.1 Definition

Frequent Pattern Mining (AKA Association Rule Mining) is an analytical process that finds frequent patterns, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other data repositories. Given a set of transactions, this process aims to find the rules that enable us to predict the occurrence of a specific item based on the occurrence of other items in the transaction.

6.2 Used Model

Apriori algorithm is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

7 Conclusion

In this paper, we have presented a yearly income prediction system Firstly we went through the Exploratory Data Analysis where we discovered our data then Data pre-processing where we went through the steps to make sure our data is ready to be implemented to give the best results followed by Classification where we went through two examples and compared their results we did the same also in the Clustering part and lastly we gave a brief Frequent pattern mining description.

References

1. ttps://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/
2. ttps://www.kdnuggets.com/2019/09/hierarchical-clustering.html
3. ttps://www.dataversity.net/frequent-pattern-mining-association-support-business-analysis/
4. ttps://towardsdatascience.com/grid-search-for-hyperparameter-tuning-9f63945e8fec
5. ttps://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623
6. ttps://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
7. ttps://www.geeksforgeeks.org/apriori-algorithm/