Introduction to Data Science
Report Assignment

Youssef Gharbi
Neptun code: IKYCUE
joe.gharbi@gmail.com
2021-12-10

ELTE
EÖTVÖS LORÁND
TUDOMÁNYEGYETEM

Eötvös Loránd University, Faculty of Informatics
Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
+36 1 372 2500
itcs@ludens.elte.hu
https://www.inf.elte.hu

Abstract. This report is the result of a long analysis of a data-set named adult provided by the US Census bureau's regional offices that are responsible for data collection, data dissemination, and geographic operations. The data-set is collected by Ronny Kohavi and Barry Becker in 1994. First part consists of the the data exploration and analysis followed by data preparation and processing. later on, the model building and data prediction using two approaches logistic regression and random forest.Then clustering is done using the K-means methodology. Finishing by frequent pattern mining using fp-growth algorithm.

1 Introduction

During this report,an analysis is conducted on the data provided by the US Census bureau's regional offices. The aim is to build a binary classifier that can predict whether the income of the person will exceed 50 thousands dollars a year or not. After that a K-means clustering algorithm is applied to the data-set to determine the likely groups of clusters. To finish, a frequent pattern mining methodology is conducted to identify the pattern features among those making more than 50 thousands a year.

However, before starting with the prediction and to have a good data-set that can be correctly fitted to the model, there will be first some data preprocessing and data exploration.

2 Exploratory data analysis

2.1 Data exploration

Exploring the data set at first is always the step needed for a good data analytic report since it gives an overview and the important consideration the reporter should take into consideration.

The data-set consists of 32561 rows with 15 features including the target feature with the following description:

1. age: (continuous, positive integer) The age of the individual.
2. workclass: (categorical, 9 distinct values) Simplified employment status of an individual
3. fnlwgt: (continuous, positive integer) Final weight of the record. Basically interpret as the number of people represented by this row.
4. education-num: (categorical, 13 distinct values) The education level, in ascending positive integer value.
5. education: (categorical, 13 distinct values) The education level. Note that for simplicity, we will ignore this column because of the existence of education-num column.
6. marital-status: (categorical, 7 distinct values) Marital status of a person.
7. occupatioin: (categorical, 15 distinct values) Rough category of the occupation.
8. relationship: (categorical, 6 distinct values) Relationship in terms of the family. Note that we ignore this column since the semantic is somewhat covered by marital-status and gender.
9. race: (categorical, 5 distinct values) Race of the person.
10. gender: (boolean) gender at-birth.
11. capital-gain: (continuous) Dollar gain of capital.
12. capital-loss: (continuous) Dollar loss of capital.
13. hours-per-week: (continous positive integer) Working hours per week.
14. native-country: (categorical, 41 distinct values) Country at birth.
15. income-bracket: (boolean) True if >= 50K, otherwise False (<50K per year).

Although the data-set is large the next figure shows the huge imbalance it has.

Almost 75% of the persons are gaining less than 50 thousands that's why this fact needs to be considered during the analysis work.

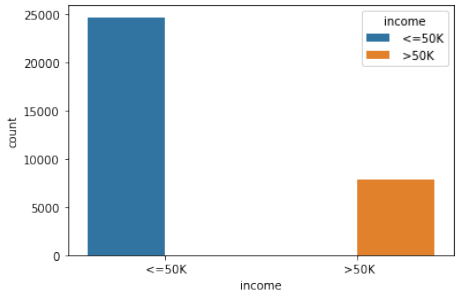

Fig. 1: Date-set distribution

Correlation:

A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. The values of the first dimension appear as the rows of the table while of the second dimension as a column. The following figure shows the correlation between the different data-set features and columns.

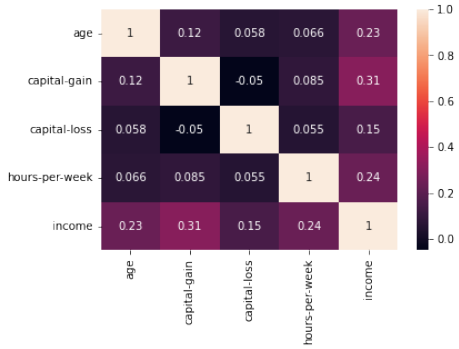

Fig. 2: Caption

The heat-map shows that there is no high correlation between the features and

columns just a slight one only between capital-gain and income. This is absolutely taken into consideration in further sections.

Features distribution:
An important step to get a better insight about the data is to visualise it's features distribution. This step gives an important overall understanding on how the data is distributed or if there is any outliers or even if the data is skewed. The following figure is showing how our numerical features are distributed.

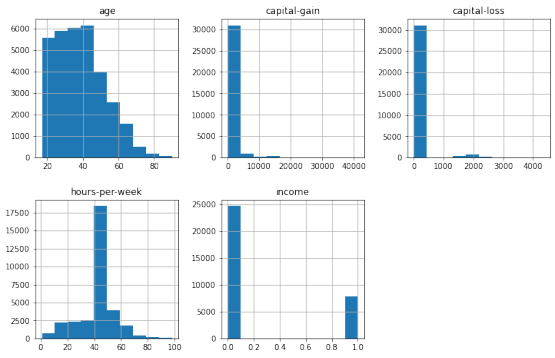

Fig. 3: Features distribution

The figure is showing that the features are not well balanced and this might effect the model and the results later that's why it has been taken into consideration while building the model.

2.2 Data cleaning

Null values:
just before feeding the model cleaning the data and dealing with missing data or noise and applying some feature engineering is important to have a good prediction and a stable model. Here the null values has been assigned the '?' mark. Thus, it require to transfer the exclamation mark to null values and then apply the correspondent method. In the case of this data-set the null values occurred in 3 features : workclass , occupation and native-country. Since the null values together present 5% of the whole data-set some would just drop it but I choose to fill it's values with mode method which assign the most frequent value to the null since it has weight on the column itself.

Checking for noisy data and outliers:
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to

the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	1077.648844	87.303830	40.437456
std	13.640433	1.055500e+05	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370510e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Fig. 4: Data-set info

```
age               73
workclass          9
fnlwgt         21648
education         16
education-num     16
marital-status     7
occupation        15
relationship       6
race               5
sex                2
capital-gain     119
capital-loss      92
hours-per-week    94
native-country    42
income             2
dtype: int64
```

Fig. 5: checking the fnlwgt outlier

As the info showed we can see that the capital-gain and hours-per-week has the value max 99999 and 99 respectively and this is due to a miss-type or an error while filling the data in. Thus, it will be fitted with mean of the columns. However, noticeable the fnlwgt is not clear yet. The next figure shows that the fnlwgt is 21648 unique values which is almost unique to the user and with further investigations in the data-set description provided by the creators and considering the result of the 3rd figure it can be considered as noisy data and as a result it will be dropped.

Feature engineering:
From figure 3 the education and education-num features has the same values one numeric and the other as written. One fair approach, is to drop one column, however what it has been considered is doing some feature engineering on those features such as material-status too will be explained in the next section.

- For education from preschool and 1st grade till 12Th grade it has been assigned to the school. HS-grade is high-school. Higher education is assigned to assoc-voc till some college.Others got undergrade, grad and doc.
- For material-status it got 3 values married not-married and other.
- Target data is encoded to binary values one and zero. One if more than 50 thousands dollars and the opposite is zero.

3 Building the model

3.1 Introduction

In this part of the report, the aim is to build and compare two machine learning models first is a logistic regression second is the rain forest model both of these models are binary classifications which predicts whether the person will have more or less than 50 thousands income a year. After cleaning the data and exploring it here comes the step before feeding the model with the data. First step it to divide the features between target data called y in this case is the

income binary form to be predicted and the other data-set features called X. Right after dividing the data-set a necessary step is to encode the categorical values since the models can only work with numerical values, there exists different methods such as OneHotEncoder ,label encoder and get dummies. In this report the method chosen is the label encoding. In this technique, each label is assigned a unique integer based on alphabetical ordering.

3.2 Logistic regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. The fundamental equation of the model is:

$$g(E(y)) = \alpha + \beta(x1) + \gamma(x2) \tag{1}$$

Where g() is link function, E(y) is the expectation of target variable and $\alpha + \beta(x1) + \gamma(x2)$ is the linear predictor.

After feeding the model with data and to check if the model is good at prediction, here comes the accuracy checking. Accuracy is the proportion of correct predictions over total predictions.

$$accuracy = correct\_predictions/total\_predictions$$

The model's accuracy result is as the following: Acc on training data: 0.802 Acc on test data: 0.798.

3.3 Random forest:

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.
We can understand the working of Random Forest algorithm with the help of following steps:

1. Start with the selection of random samples from a given dataset.
2. Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
3. In this step, voting will be performed for every predicted result.
4. At last, select the most voted prediction result as the final prediction result.

After implementing the random forest algorithm the results on training and testing data are as the following: Acc on training data: 0.802 Acc on test data: 0.798.

3.4 checking the results

Confession matrix:
Confusion Matrix is a performance measurement for machine learning classification. It is a table with 4 different combinations of predicted and actual values (TP, FP, FN, TN).

- True Positive(TP): predicted positive and it's true.
- True Negative(TN): predicted negative and it's true.
- False Positive(FP, Type 1 Error): predicted positive and it's false.
- False Negative(FN, Type 2 Error): predicted negative and it's false.

TP: 6743	FP: 654
FN: 907	TN: 1465

Table 1: Confession matrix for logistic regression

TP: 6912	FP: 485
FN: 1487	TN: 885

Table 2: Confession matrix for random forest

Classification report:
Before giving the classification report for both models here what the report has:

- Recall=TP/(TP+FN).
- Precision=TP/(TP+FP).
- F-measure=2*Recall*Precision/(Recall+Precision).
- The support is the number of occurrences of each class in y_true.

	precision	recall	f1-score	support
0	0.82	0.93	0.88	7397
1	0.65	0.37	0.47	2372
accuracy			0.80	9769
macro avg	0.73	0.65	0.67	9769
weighted avg	0.78	0.80	0.78	9769

Fig. 6: Classification report for Logistic regression

	precision	recall	f1-score	support
0	0.88	0.91	0.90	7397
1	0.69	0.62	0.65	2372
accuracy			0.84	9769
macro avg	0.79	0.76	0.77	9769
weighted avg	0.84	0.84	0.84	9769

Fig. 7: Classification report for Random forest

As the matrices and the classification reports shows the logistic regression is less accurate than the random forest however it is faster while the random forest

takes longer to execute yet it gives better results. These results are relative to this model only, meaning if the fitted data or the hyper-parameters are different it will give different results.

4 Clustering

Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger data-sets without concern for the specific outcome. Before feeding the clustering algorithm a slight data tuning is needed. First PCA dimension reduction is needed. It mainly refers to reducing the number of input variables for a data-set. So in this data-set the input is 14 features and the result of the reduction is 2 features. This is possible with sklearn decomposition library. Next, MinMax scaling is needed. It transforms features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

$$X\_std = (X - X.min(axis = 0))/(X.max(axis = 0) - X.min(axis = 0))$$

$$X\_scaled = X\_std * (max - min) + min$$

where min, max = feature_range.

choosing the optimal cluster number with elbow method:
A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

1. Distortion: It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.
2. Inertia: It is the sum of squared distances of samples to their closest cluster center.

Based on elbow method figure 8 the optimal number of clusters is 2.

K-means on the data-set The basic steps for the k-means clustering:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).

6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

After running the k-means algorithm with the data and setting the number of clusters to two as computed by the elbow method the result is in the figure 9.

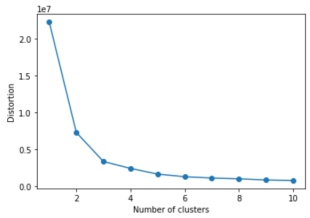

Fig. 8: Optimal number of clusters

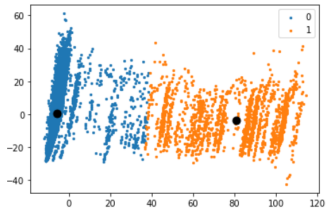

Fig. 9: Clustering of the data-set

5 Frequent pattern mining

Frequent Pattern Mining is an analytical process that finds frequent patterns, associations, or causal structures. The algorithm used is provided by mlxtend library named association rules.

	support	itemsets
0	0.913762	(United-States)
1	0.854274	(White)
2	0.669205	(Male)
3	0.753417	(Private)
4	0.798716	(United-States, White)
5	0.611406	(United-States, Male)
6	0.682749	(United-States, Private)
7	0.642302	(White, Private)

Fig. 10: Association rules on the data-set

Typically, support is used to measure the abundance or frequency of an itemset in a database. An itemset is "frequent itemset" if the support is larger than a specified minimum-support threshold.An association Rule is an implication expression of the form X ->Y, where X and Y are any 2 itemsets.
Based on the association rule result of the figure 10 a deduction is that the United states is more frequent than the others followed by the white and then male sex. Also, Being from the US and white is more likely to get more than 50 thousands especially if the work is private. Of course there can be different deduction based on different methods but this is relative to only this model.

6 Conclusion

After starting with data exploration like the correlation and features distribution, comes data cleaning from null values and noise detection and fixing and preparation and once the data is ready, it was fitted to two different classification algorithms Logistic regression and Random forest.There was some results reports and a relative comparison. Right after comes clustering the data-set and implementing a frequent pattern mining algorithm to look for some patterns in the data. These results are fully relative to the model and logic chosen into this report. The model can be improved and can have more tuning for better results.

References

1. Classification algorithms - random forest, 2021. URL: https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_random_forest.htm.
2. Sarang Narkhede. Understanding confusion matrix - towards data science, 05 2018. URL: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.
3. Koo Ping Shung. Accuracy, precision, recall or f1? - towards data science, 03 2018. URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.
4. Wikipedia Contributors. Precision and recall, 10 2021. URL: https://en.wikipedia.org/wiki/Precision_and_recall#Definition_(classification_context).
5. Ml clustering: When to use cluster analysis, when to avoid it, 02 2020. URL: https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/.
6. sklearn.preprocessing.minmaxscaler, 2021. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.
7. Elbow method for optimal value of k in kmeans - geeksforgeeks, 06 2019. URL: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.