Introduzione al Machine Learning
Project - first part

Abel Ficano

Data exploration

Data is structured in 15 numerical variables and one boolean target variable. The variables are the following:

- AGEP: age (number between 0 and 99)
- SCHL: educational attainment (1 to 24)
- MAR: marital status (1 to 5)
- RELP: relationship (1 and 17)
- DIS: disability recode (1 for yes, 2 for no)
- ESP: employment status of parents (1 to 8)
- CIT: citizenship status (1 to 5)
- MIG: mobility status (1 to 3)
- MIL: military service (1 to 4)
- ANC: ancestry recode (1 to 4)
- NATIVITY: 1 if native, 2 if foreign-born
- DEAR: hearing difficulty (1 yes, 2 no)
- DEYE: vision difficulty (1 yes, 2 no)
- DREM: cognitive difficulty (1 yes, 2 no)
- SEX (1 male, 2 female)

The target variable is:

- ESR: employment status recode

Missing values are present in some of the columns, for the moment we replace them with NaN values. To see correlations between variables, let's evaluate the correlation matrix:

	AGEP	SCHL	MAR	RELP	DIS	ESP	CIT	MIG	MIL	ANC	NATIVITY	DEAR	DEYE	DREM	SEX	ESR
AGEP	1.0	0.53	-0.65	-0.16	-0.32	-0.5	0.14	-0.09	0.61	-0.08	0.16	-0.24	-0.13	0.3	0.05	0.11
SCHL	0.53	1.0	-0.44	-0.11	-0.03	-0.53	0.01	0.05	0.71	-0.06	0.03	-0.04	-0.0	0.51	0.02	0.43
MAR	-0.65	-0.44	1.0	0.32	0.08	0.4	-0.2	0.09	-0.49	0.07	-0.21	0.1	0.03	-0.24	-0.04	-0.24
RELP	-0.16	-0.11	0.32	1.0	-0.09	0.03	-0.01	0.2	-0.03	0.05	-0.02	-0.0	-0.04	-0.09	-0.04	-0.14
DIS	-0.32	-0.03	0.08	-0.09	1.0	0.09	0.04	-0.0	-0.1	-0.01	0.03	0.51	0.41	0.19	-0.0	0.21
ESP	-0.5	-0.53	0.4	0.03	0.09	1.0	-0.14	-0.02	-0.65	0.04	-0.15	0.06	0.04	-0.29	-0.01	-0.32
CIT	0.14	0.01	-0.2	-0.01	0.04	-0.14	1.0	-0.01	0.21	-0.15	0.98	0.03	-0.0	0.11	0.01	0.11
MIG	-0.09	0.05	0.09	0.2	-0.0	-0.02	-0.01	1.0	0.04	0.03	-0.02	0.02	-0.0	0.03	-0.01	0.02
MIL	0.61	0.71	-0.49	-0.03	-0.1	-0.65	0.21	0.04	1.0	-0.07	0.22	-0.04	-0.05	0.39	0.08	0.46
ANC	-0.08	-0.06	0.07	0.05	-0.01	0.04	-0.15	0.03	-0.07	1.0	-0.16	0.0	0.0	-0.05	0.0	-0.02
NATIVITY	0.16	0.03	-0.21	-0.02	0.03	-0.15	0.98	-0.02	0.22	-0.16	1.0	0.02	-0.0	0.11	0.02	0.11
DEAR	-0.24	-0.04	0.1	-0.0	0.51	0.06	0.03	0.02	-0.04	0.0	0.02	1.0	0.23	0.05	0.03	0.1
DEYE	-0.13	-0.0	0.03	-0.04	0.41	0.04	-0.0	-0.0	-0.05	0.0	-0.0	0.23	1.0	0.07	-0.01	0.07
DREM	0.3	0.51	-0.24	-0.09	0.19	-0.29	0.11	0.03	0.39	-0.05	0.11	0.05	0.07	1.0	0.01	0.27
SEX	0.05	0.02	-0.04	-0.04	-0.0	-0.01	0.01	-0.01	0.08	0.0	0.02	0.03	-0.01	0.01	1.0	-0.07
ESR	0.11	0.43	-0.24	-0.14	0.21	-0.32	0.11	0.02	0.46	-0.02	0.11	0.1	0.07	0.27	-0.07	1.0

We immediately see some (not so large) obvious correlations, for example between age and education or between disability and hearing difficulty.

However, the correlations we are most interested in are those regarding employment status: we see some correlation with military service and educational attainment, and some smaller correlation with cognitive difficulty and disability. Even though we would expect a certain correlation with age, it isn't very high.

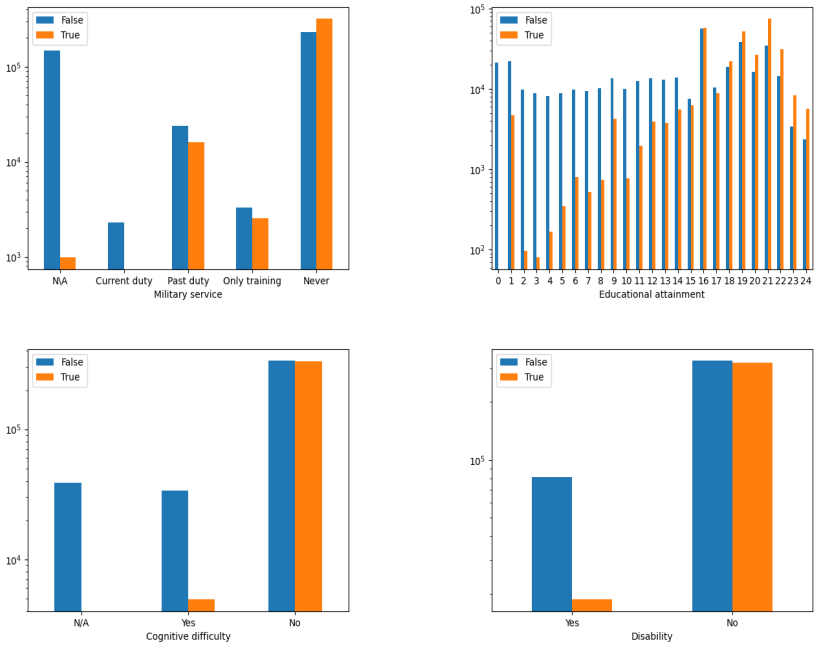

Figure 1: Bar graphs of correlated variables

To further investigate the detected correlations, we show the bar graphs obtained for each variable of interest after splitting the dataset into two parts, based on the target variable (Fig. 1).

In light of the bar graphs, we can now correctly interpret the found correlations: people without data on military service or currently on active duty are almost always unemployed, as well as children in lower grades and people with cognitive difficulty or disability.

These associations, which we found visually through bar graphs, could be more thoroughly investigated through association analysis algorithms. Instead, we will now delve into unsupervised clustering techniques.

Clustering

We already know we want to find two clusters in the data, so the best algorithm for this case is K-means. First of all, we have to deal with missing values. We choose to assign 0 values to the missing values: in this way, we are considering unknown values as an additional category. Then data is preprocessed through a MinMaxScaler normalization.

Now we can run the K-means algorithm: we choose as hyperparameters 50 random initial conditions and 1000 maximum iterations. The algorithm separates data into two clusters of sizes 150,755 and 595,581, while the target value separates data into clusters of sizes 338,116 and 408,220.

Evaluation

To evaluate the separation of the clusters we use the Davies-Bouldin score. For our data, it is 1.47, which is quite high and means that the clusters are not well separated.

For external evaluation, i.e. evaluation against the ground truth, given in this case by the values of the target variable, we use the Adjusted Rand Index, which has a value of 0.088. It is low, meaning that label assignment is not much better than random assignment. This is also confirmed by the percentage of correct labels, which is 65%.

Also dropping the missing values performance is not satisfactory, even though a little better: Davies-Bouldin score is 1.45, ARI is 0.089, and the percentage of correct labels is 65%.

Bad performance is expected in this case because the data encoding used is inappropriate for clustering: even though the data is given in numeric format, almost all attributes are categorical and do not have a favourable ordering.

More specifically, while age and educational attainment have an ordering so numerical encoding is natural, for other attributes, like military service, every permutation of the used encoding would be equally right.

To improve performance, we could try permuting the numeric codes used for the categorical attributes or flattening out the categorical attributes to have

only numerical and numerically encoded boolean attributes. But the best choice is surely to turn to supervised learning.

Classification

Decision tree

The first supervised learning model we try to apply is a decision tree. We split the dataset into a test (30%) set and a training set (70%) and trained a decision tree on the training set, using the Gini index as the measure of node impurity.

Running the decision tree on the training set gives an accuracy of 89% and the following normalized confusion matrix:

	Predicted positive	Predicted negative
Positive	89%	11%
Negative	11%	89%

However, the true indicator of a model's validity is its performance on the unseen test set. In this case, our decision tree gives a 78% accuracy and the following normalized confusion matrix:

	Predicted positive	Predicted negative
Positive	77%	23%
Negative	21%	79%

The performance of the model is good both on the training set and on the test set. To evaluate the complexity of the model we can consider its depth and the number of leaves. Our decision tree has a depth of 42 and 80,954 leaves. Given that the number of variables was high (fifteen) the tree's complexity is reasonable.

SVM

Let's try another supervised learning model for classification, support vector machine.

Since we have a large data set, using directly the SVC implementation of SVM would be too time-consuming. So, instead, we first apply a nonlinear transformation to data (specifically a Nystroem transformer) that acts as a kernel approximation and then train a linear version of SVC on the transformed data.

This gives us a 79% accuracy and the following normalized confusion matrix both on the training data and on the test data:

	Predicted positive	Predicted negative
Positive	80%	20%
Negative	22%	78%

Using an SVM we obtained slightly better results compared to the decision tree. Since the model performs equally on training and test data we know for sure that the model has not overfitted.