Big Data Analytics Project : Dataset Adult from UCI

Mohammed Meftah, Raunaq Paul and Binxiang Xiang

16^{th} January 2017

Contents

1 Introduction

For this project, we have decided to work on the dataset Adult from UCI. This is a classification dataset with two classes, as we have to understand and predict whether a person earns less or more than 50k dollars a year, according to the provided features. Section 2 describes in details the used dataset Adult. Section 3 highlights how we first processed data. Section 4 shows some statistical analysis on the dataset. Section 5 presents the methodology and the performances of different algorithms (list of tried algo to add) on the prediction of our target variable.

2 Dataset description

The raw dataset has 48842 rows and 15 columns including the target variable. This data is originally from the american census bureau and shoulb be representative of the american population. This dataset is mixing continuous variables and discrete variables as follows :

age : continuous, age of the individual.

fnlwgt : continuous, the number of people the observation should represent, not useful for us.

workclass : discrete, 8 categories representing the type of the employer of the individual (Private, Federal-gov, Never-worked ...).

education : discrete, 16 categories representing the highest level of education achieved by the individual (Bachelors, Masters, Doctorate, Highschool grad ...).

educationnum : continuous, number of years of education.

mstatus : discrete, 7 categories representing the marital status of the individual(Divorced, Never-married, Separated ...).

occupation : discrete, 14 categories representing the occupation of the individual (Tech-support, Sales, Exec-managerial, Other-service ...).

relationship : discrete, 6 categories representing the relationship of the individual, we don't know with who so this variable is not very clear (Wife, Husband,Own-child ...).

race : discrete, 5 categories representing the race of the individual (White, Asian-Pac-Islander, Black, Amer-Indian-Eskimo and Other).

sex : discrete, sex of the individual (Female or Male).

capitalgain : continuous, represents the recorded capital gains.

capitalloss : continuous, represents the recorded capital losses.

hoursperweek : continuous, represents the number of hours worked per week.

nativecountry : discrete, 41 categories representing the country of origin of the individual.

target : discrete, variable to predict, 2 classes ($<= 50k\$$ or $> 50k\$$).

For the next sections, these names will be used to refer to the corresponding variable.

3 Data processing

In this section, we explain the different steps of processing we did on our raw dataset.

3.1 Basic processing :

We did perform some basic processing as a first step :

- Remove the variable *fnlwgt*, as it is not an information on the individual.
- Remove the variable *educationnum*, as it contains the same information as the variable *education* and we would prefer to work with a discrete variable.
- Transform the target variable into binary values, 0 if $<= 50k\$$ 1 if $> 50k\$$.
- Remove rows with missing values, indeed variables *workclass*, *occupation* and *nativecountry* don't always have values. It represents 3620 rows.
- Scale continuous variables : *capitalgain,capitalloss and hoursperweek*

After these steps, the new dataset has 45222 rows. And 24.8% of the observations are in the class 1 and 75.2% are in the class 0.

3.2 Advanced processing :

As the discrete variable *nativecountry* contains 41 categories where the main one (United States) represents 91.3% of the total, we have decided to group countries together into bigger categories. We don't change United-States, Canada, Japan, China, India , Iran and Mexico. The changes are the following :

New value	Old values concerned
WEurope	England, France, Germany, Holand-Netherlands, Ireland, Italy, Portugal, Scotland
EEurope	Greece, Hungary, South, Yugoslavia
SEAsia	Cambodia, Laos, Philippines, Thailand, Vietnam
LatAmerica	Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, Guatemala, Haiti, Honduras, Jamaica, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Puerto-Rico, Trinadad&Tobago
China	HongKong, Taiwan

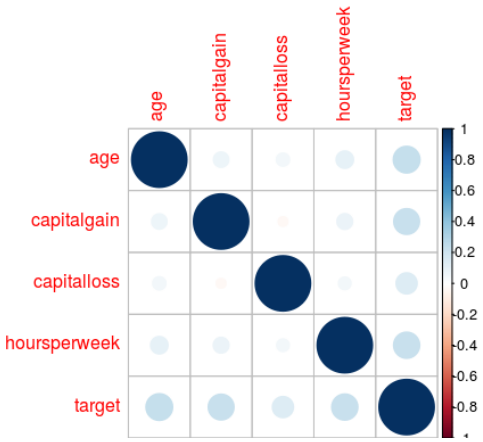
After the transformation, our *nativecountry* variable has now 11 different values against 41 before. Of course, we lost some information but it shoud not be so significant as we regroup countries from same area and with similar economic situation.

4 Analysis

After the processing part, we can start to look at data and make some analysis by variable and by class. We separated the study whether the variables are numeric or categorical.

4.1 Numerical features

We first study the correlation of numerical variables with the target. We thus made the following corrplot :



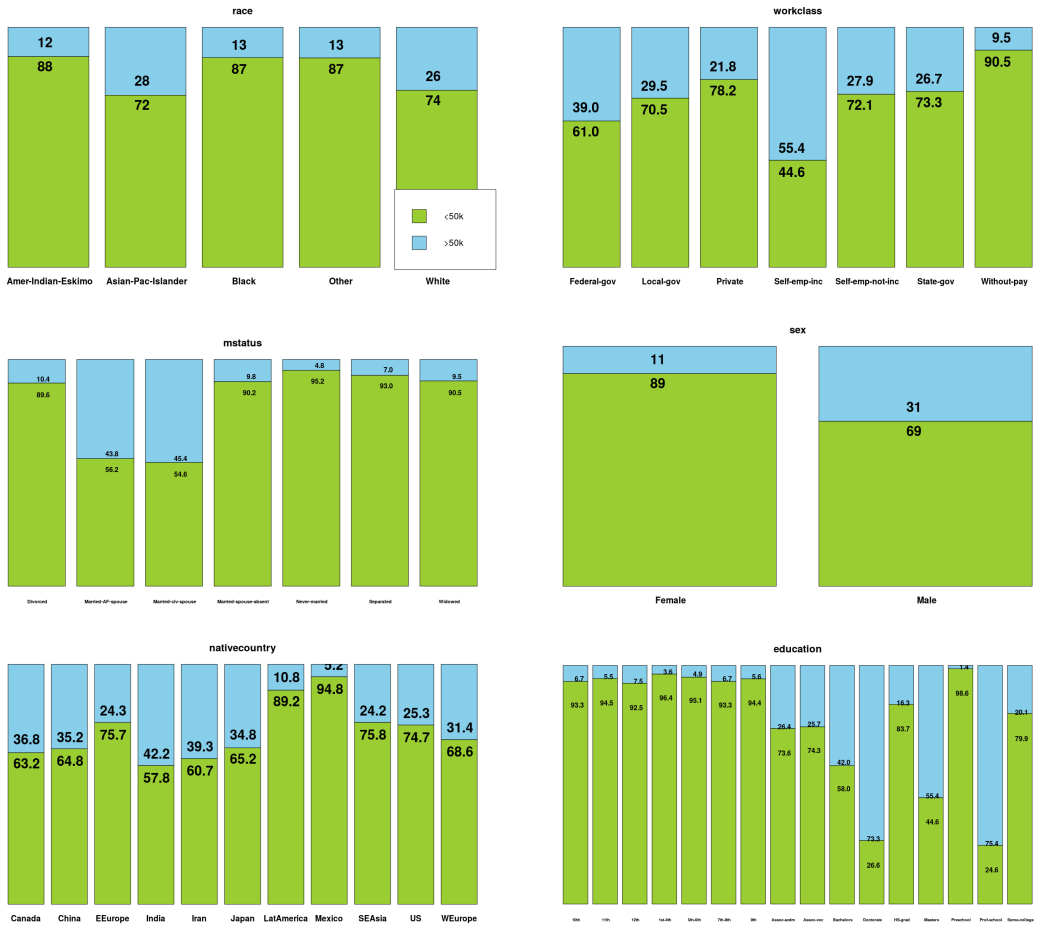
It shows that all the numerical variables are correlated positively with the target. However the correlation values remains low with the best score for (age, target) around 0.3. So for a first model, maybe it is not of paramount importance to include such variables.

4.2 Categorical features

We cannot define a correlation properly with categorical variables. We choose to look at the distribution of the target within the categorical variables and test their independance (χ-squared test) with the target in order to have an idea of each possible contribution.

Feature	χ-square	p-value
race	452.3	<2.2e-16
workclass	1207.3	<2.2e-16
mstatus	9109.2	<2.2e-16
sex	2104.1	<2.2e-16
nativecountry	379.36	<2.2e-16
education	6000	<2e-16

The χ-square tests are very concluding so we are confident to reject the null hypothesis and consider that each of these categorical variables are non-independant with the target, it means that they are correlated. The very low p-values pinpoint also this high correlation. We should see some pattern by looking at the target distribution for each unique element of each feature.



As we clearly see, for almost all this categorical features, the distribution of the target is not homogenous within the feature labels. It should enable us to make better predictions with these features as a basis. And some of the distributions are confirming

some common knowledge. For example, in proportion more men (31%) are in the $> 50k\$$ category than women (11%) or people with PhD (73%) are more prone to be in the class 1 than people with master degree (53.4%) or bachelor degree (42%). We can see some other interesting insights too, for example, divorced people (10.4%) are less prone to be in the class 1 than married people (≈44%). In terms of proportion, more Asian people (28%) are earning more than 50k$ than other minorities(12-13%) and they even outreach white people.

5 Methodology and results

5.1 Methodology

In this section, we'll describe the procedure in order to compare different algorithms we have chosen on our dataset. These algorithms are Logistic Regression (LogReg), Bagging, Random Forest (RF) and K-Nearest Neighbours. The difficulty here is that our dataset is mixing continuous and discrete variables. For the first three methods, it wasn't an issue as they deal easily with a mix of continuous and discrete variables. But for K-Nearest Neighbours, we had to transform the discrete variables before applying the model. We'll describe very precisely what we did for KNN and we'll see that the performances are surprisingly very good for KNN.

To compare the different algorithms, we perform the following procedure :

1. Find the best hyperparameters for our model by Cross Validation
2. Split randomly dataset into train and test sets, 70% in the train set and 30% in test set.
3. Train the algorithm on the train set
4. Predict values on the test set
5. Save the prediction error with accuracy measure
6. Repeat 20 times steps 2 to 5
7. Do steps 1 to 6 for each method (algorithm)

5.2 Results

5.2.1 Logistic Regression, RF and Tree Bagging

We'll not detail the hyperparameters selection for Random Forest and Tree Bagging. Here are the results of Logistic Regression, optimized Random Forest and optimized Tree Bagging :

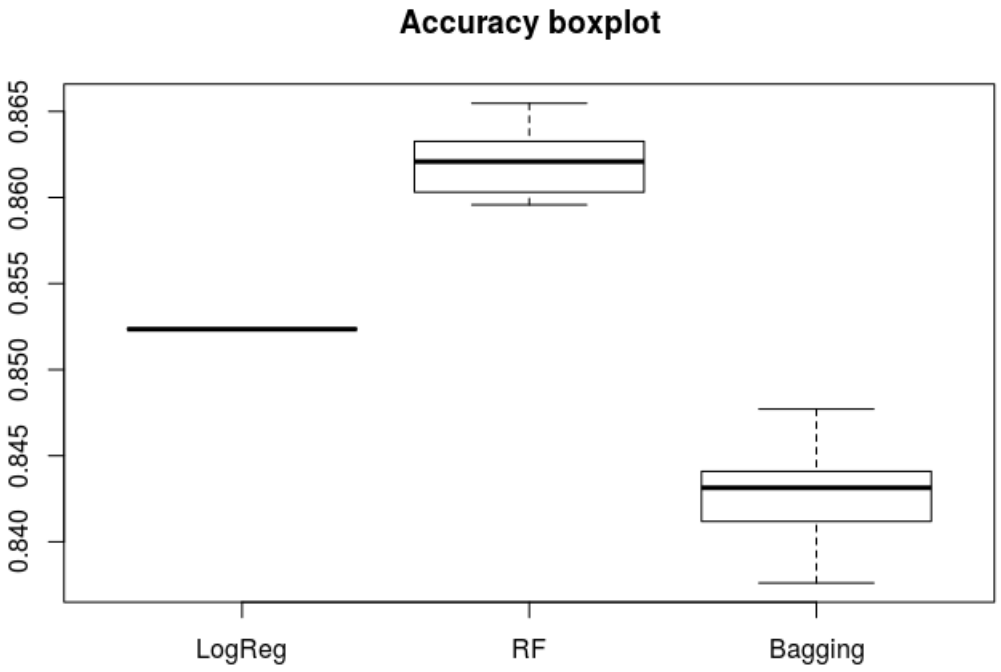


Figure 1: 20 runs Accuracy boxplot for LogReg, RF and Tree Bagging

Among these three methods, the best performing one is Random Forest with an average accuracy of **0.862**. Furthermore, we can notice that Logistic regression is very robust. The performances are very stable even though the train and test set are not the same. We already knew that tree bagging will perform worse since RF are an enhancement of this very method.

5.2.2 K-Nearest Neighbours

As we said before, we can't directly apply KNN on our dataset. We need to transform our discrete variables into numeric variables, on which KNN will be able to build a metric. Besides, we can't just create a correspondence table. Indeed, if we choose to

code the n values/categories of a discrete variable by a set of n integers, we'll introduce some order relationships. For instance, for the variable race, if we code 'White' by 1, 'Black' by 2 , 'Asian' by 3 and so on, we will have relationships like 'Black' > "White' or 'Asian' = 3*'White', which are not true here.

A common method to solve this problem is to transform discrete variables into dummy variables. That's what we did on our dataset, we transformed each discrete variable into dummy variables. At the end, we obtained 72 features excluding the target variable. The number of features is too big, we need to perform dimension reduction either because of the curse of dimensionality or because of the computational time. However, the screeplot (Fig. 2) of a PCA on our dataset did not give us something very remarkable so we decided to choose the number of features to keep by looking at the performances.

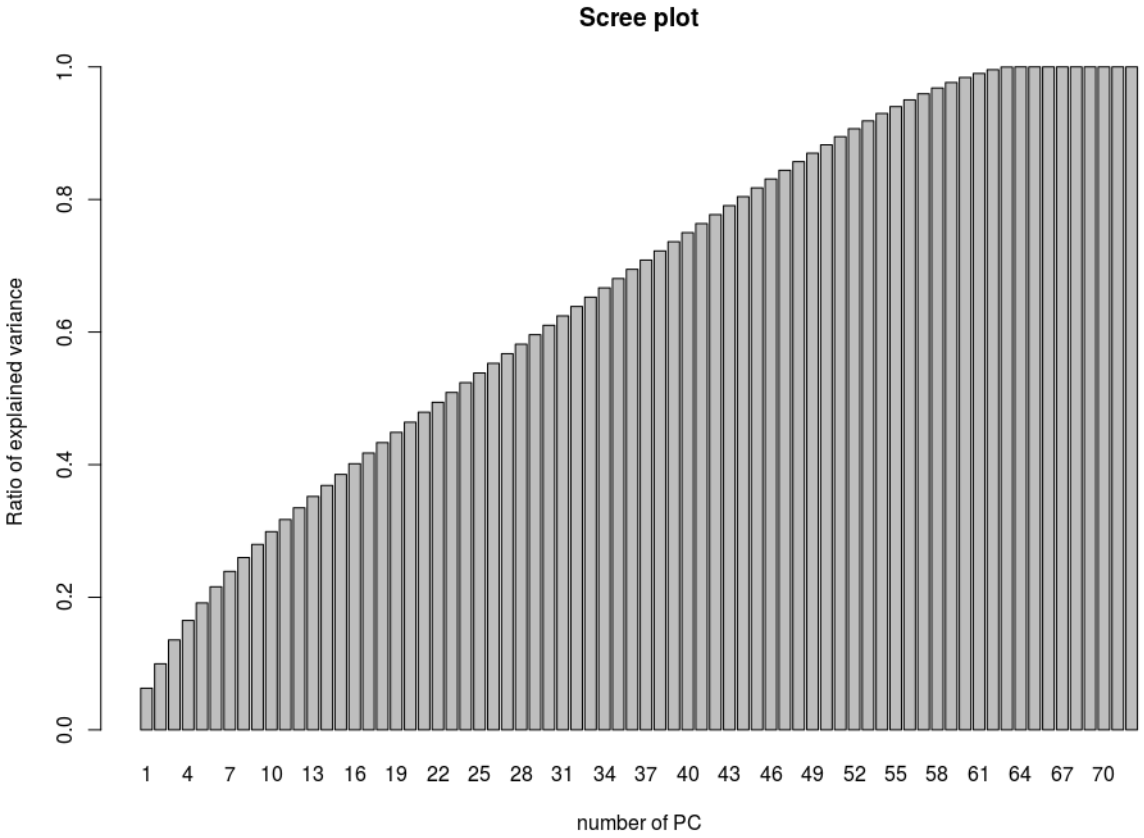


Figure 2: Screeplot for our dataset

Figure 3 presents the boxplot of a 3-NN for different number of kept features (after 5 runs):

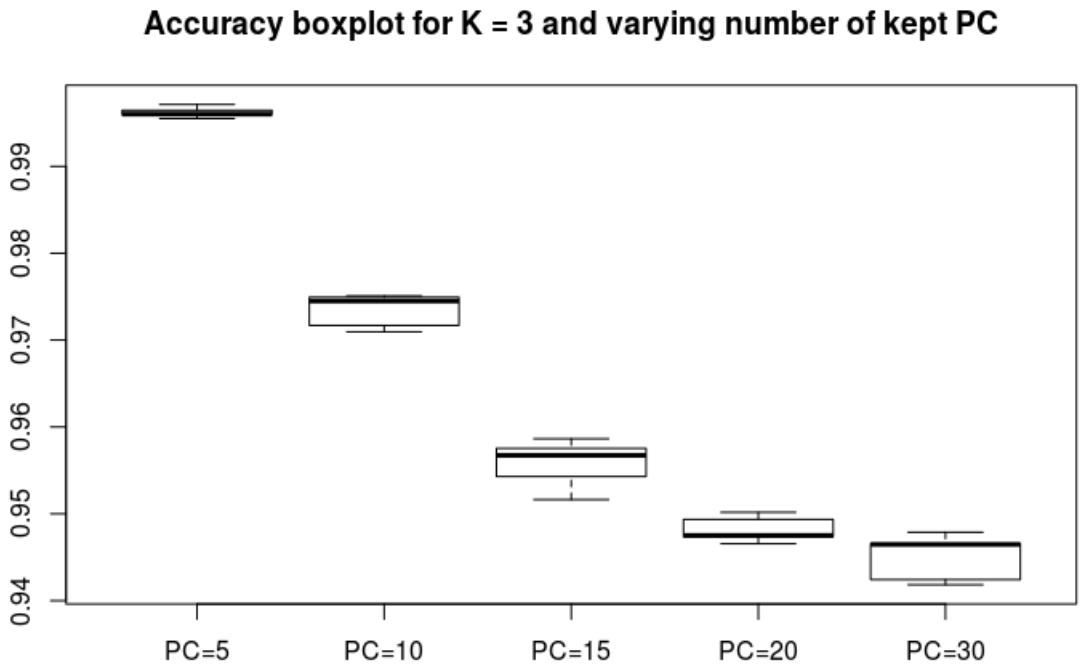


Figure 3: Accuracy boxplot for 3 NN and different number of principal components

So, we can see that if we project our data into a space of dimension 5 with PCA, the performances are astonishingly good (average accuracy of even if the corresponding explained variance is around 20%. This could be explain that given the fact that the categorical variables are highly correlated to the target, it is unnecessary to complexify the models with a lot of variables (otherwise themodels could be subject to the curse of dimensionality). Without doubt, we decide to keep the 5 first principal components in order to project our data.

So with this parameter fixed, figure 4 presents performances when we vary the parameter K of number of nearest neighbours.

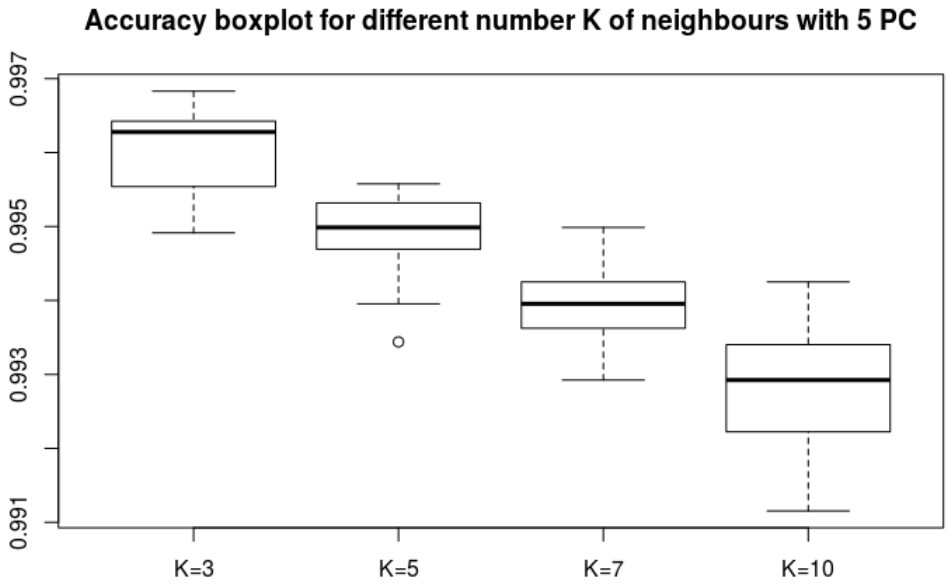


Figure 4: Accuracy boxplot for varying number of Nearest Neighbours and 5 PC

So, performances are the best for K=3. We have an average accuracy of **0.996**. This method outperformed Logistic Regression, Random Forest and Bagging. But the problem is that we can't really explain which features were valuable as we performed PCA on our dataset.

6 Conclusion

The results we obtained were surprisingly good. Given the particularities of the dataset, namely the intrication of numerical and categorical variables, it required quite a lot of processing and manual adjustment to get to these results. This kind of study highlights perfectly the role of the data scientist for a real-world dataset that demands more than a "click-and-go" algorithm. In particular, we processed the data to make it understandable for the different methods (cleansing, formatting, dummy variables etc..), we then performed a data analysis to make sure to take into accoount in our models the features that could explain the target. It allows us to justify the choice of "common sense" explanations. We then train and test several methods on different parts of the data to make sure the models are not overfitting. This methodoly allows to get outstand-

ing results on this particular dataset. Of course the relative small size of the dataset made this tractable for our computational power. However this methodology, involving dimensionality reduction, is scalable to larger dataset.