Adult Census Income

Julia Bui Xuan[1], Andreea Maria Dobre[1], Michele Salvaterra[1], Luca Sammarini[1], Eugenio Tarolli Bramè[1]

Abstract

This project aims to predict whether a person earns more than 50k/year, taking into account some socio-economic input attributes. Supervised machine learning methods are applied in order to solve the classification task and different models are performed. The analysis deals with the class imbalance problem, handled with the minority class undersampling. Trying to understand the most valuables attributes, feature selection is implemented using both multivariate filter and wrapper. J48, Random Forest and Logistic Regression appear to be the most appropriate models. The performance is evaluated in terms of Accuracy, Precision, Recall, F_1 measure and ROC-Curve.

Keywords

Classification models — Cross Validation — Class Imbalance — Feature Selection

[1] *CDLM in Data Science, Università degli Studi di Milano-Bicocca*

Contents

1. Introduction

Annual income refers to the total money earned over a year, before taxes. It includes salary, tips, commissions, overtime and bonuses accrued over the year.

Information about a person's income can be relevant in several domains, i.e. marketing, insurance-based, banking and also taxation level.

For instance, it can have a great influence on the amount of taxes to be paid. This information can be used by governments to predict the total amount of money available for the government spending. Furthermore, it can be hugely significant also at the business level. For instance, businesses which target individuals with high income can use Machine Learning models in order to approach only people whose financial situation matches specific requirements. This can lead the businesses to increase the rate between sold services and approached people, saving money and time.

In addiction, people generally are not willing to share their personal information about the financial state. For this reason, income is a feature that is difficult to discover.

Therefore, the purpose of this project is to find a classification model able to predict the income based on some easily accessible socio-economic attributes.

2. Dataset Structure

The dataset used to develop the project was found on Kaggle [1] and was extracted from the 1994 Census Bureau Database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The last update of the database dates back to 2016/10/08.

The dataset is made of 32561 observations and consists of 15 attributes, split into 4 numeric variables and

11 categorical variables.

The attributes present in the database are the following:

1. ***age***, only people above 17 years old and under 90 years old are considered;
2. ***workclass***, describes in which sector a person works;
3. ***fnlwgt***, a final weight which gives similar values to people sharing similar socio-economic characteristics;
4. ***education***, indicates the degree of education of a person;
5. ***education.num***, indicates the degree of education of a person expressed by a number;
6. ***marital.status***, indicates whether a person is single, married, separated, divorced or widowed;
7. ***occupation***, indicates the profession of a person;
8. ***relationship***, indicates the status of a person's relationship;
9. ***race***, characteristic of a person depending on their physical features;
10. ***sex***, category in which a person falls on the basis of their reproductive functions;
11. ***capital.gain***, refers to the increase of a person's capital;
12. ***capital.loss***, refers to the decrease of a person's capital;
13. ***hours.per.week,*** refers to the work hours in a week;
14. ***native.country***, refers to the country a person was born;
15. ***income***, refers to the money earned per year. It is considered as the target variable. It can take two values:
    - $\leq 50K$
    - $> 50K$

The attributes *education* and *education.num* express the same information in different ways. In facts, there is a complete correspondence between their categories. For this reason, it has been decided to keep only the latter for the study.

The number of observations with income less than or equal to 50K/year is 22654. On the other hand, the number of observations with income greater than 50K/year is 7508. Therefore, the latter will be considered as the positive (rare) one. The dataset is unbalanced with respect to the target variable.

2.1 Statistical hints

In this section it is possible to observe how the dataset is distributed according to some attributes.

Sex	Marital Status	Age
Male 68%	Married 46%	17-37 50%
Female 32%	Never-married 33%	37-48 25%
	Other 21%	48-90 25%

Workclass	Native Country	Race
Private 70%	U.S.A 90%	White 85%
Self-emp-not-inc 8%	Mexico 2%	Black 10%
Other 22%	Other 8%	Other 5%

A correlation analysis was carried out between the four numerical attributes present in the dataset; a non-correlation was obtained between them.

3. Methodologies

3.1 Missing removals

The number of missing values of the considered dataset is 4262, belonging to 2399 observations. Missing values in a dataset refer to the incompleteness in attribute value. This may occur for different reasons, such as measurement error, ignorance, data corruption and equipment failure and can lead to some challenges when analysing data. Since the observations containing missing values represent only 7% of the dataset, record removal procedure was implemented, decreasing their number to 30162.

3.2 Cross validation

The performance of a classification model depends on the technique used to divide the dataset into train and test set. The k-fold cross validation technique was

applied, with k=5. It has a smaller bias compared to the standard holdout and the iterated holdout procedures, because it is able to reduce the impact of outliers. Furthermore, it guarantees that every record of the dataset is included in the training set the same number of times and in the test set exactly once. The dataset is partitioned into k-folds (k exhaustive and mutually exclusive subsets), containing the same number of records. K iterations are performed using a different fold as test set at each iteration. Finally, the performance measures are obtained by computing the average of all the measures computed during the k-iterations.

3.3 Classification models

In this project different classification techniques have been implemented with the objective to find the most appropriate one. Six classification models were performed, belonging to four macro categories:

- **Heuristic models**: following a tree-based approach, these algorithms allow to build interpretable solutions. In particular, decision tree J48 and Random Forest are used;
- **Regression based-models**: logistic regression is used to solve the binary classification task. It is applicable to continuous attributes and with certain accuracy also to nominal attributes;
- **Separation models**: Support Vector Machine (SMO in Weka) and Multi-layer Perceptron models are used;
- **Probabilistic models**: the Naive Bayes model is also evaluated.

3.4 Performance Measures

In order to compare the performance of the models implemented, different measures were used. In particular accuracy, precision, recall, F_1 measure and AUC were computed.

Accuracy measures the model's ability to give reliable classifications on new records and it is defined as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{1}$$

It is normally reliable when classes are equally balanced with respect to the target variable. However, an imbalanced dataset is analysed. For this reason, different measures are needed when one class is less frequent (positive class) than the other (negative class). These measures are the following: ***Precision***, ***Recall*** and $\boldsymbol{F_1}$ ***measure***.

Precision quantifies the number of positive class predictions that actually belong to the positive class.

$$p = \frac{TP}{TP+FP} \tag{2}$$

The higher the precision (p), the lower the number of false positives (FP) committed.

Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

$$r = \frac{TP}{TP+FN} \tag{3}$$

A high Recall (r) means few erroneously positive records classified as a negative class. In fact, the Recall is equivalent to the true positive rate (TPR).

F_1 measure is defined as the harmonic mean between Recall and Precision.

$$F_{measure} = \frac{2rp}{r+p} \tag{4}$$

As defined, a high value implies high recall and precision values.

The **ROC curve**, *receiver operating characteristic curve*, is a graphical plot created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It represents the performance of a classifier without considering the distribution of the class and therefore it is used to compare different models taking into account the area under the curve (AUC).

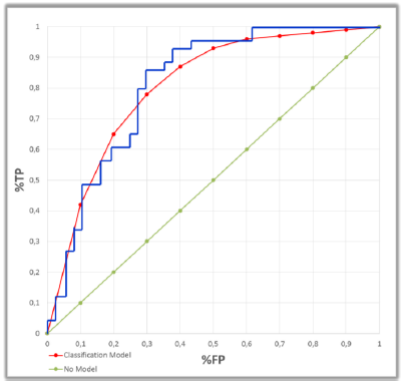

Figure 1. Example of ROC curve

3.5 Class Imbalance Problem

The class imbalance problem corresponds to domains for which one class is represented by a large number of examples while the other is represented by only a few. [2]

This can have a significant effect on the performance of the classification model: if the dataset is strongly unbalanced, the model tends to behave as the ZeroR Rule. This means the algorithm predicts the class value that has the most observations in the training dataset. In order to solve this problem a sampling based approach was used. This can be roughly classified into two categories:

- oversampling, by adding more of the minority class so it has more effect on the machine learning algorithm;
- undersampling, by removing some of the majority class so it has less effect on the machine learning algorithm.

By oversampling, just duplicating the minority classes could lead the classifier to overfitting. By undersampling, there is the risk of removing some of the majority class instances which are more representative, thus discarding useful information. [3] However, considering that the total number of observations belonging to the minority class is large, undersampling approach was used.

3.6 Feature selection

There are many situations in which using all the attributes that are present in the entire dataset is not efficient. For this reason, in order to solve the classification problem, it is very important to proceed with a feature selection aiming to discover which attributes are *redundant* and *irrelevant*. In this way, the number of input variables and the computational cost are reduced, the performance is improved and a better comprehension of the model applied can be reached.

There are several approaches to detect these attributes, for example *Brute-force*, *Embedded*, *Filter* and *Wrapper*. The latter two were used in this project.

- **Filter**: this model relies on general characteristics of the training data to select some features without involving any learning algorithm. When the number of features becomes very large, the filter model is usually chosen due to its computational efficiency.
- **Wrapper**: the attributes are selected on the basis of the classification model chosen. Those that are relevant for one model could not be relevant for another. In the wrapper approach, the feature subset selection is done using the induction algorithm as a black box. The feature subset selection algorithm conducts a search for a good subset using the induction algorithm itself as part of the evaluation function. [4]

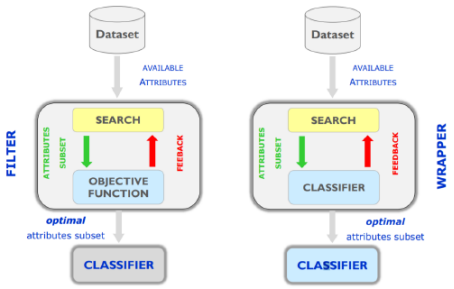

Figure 2. Filter and Wrapper approaches

4. Experimental results

The main objective of this study is to predict the income of a person, taking into account some socio-economic variables. In order to perform this task, six classification models were used and compared in different settings.

The cross validation procedure with a 5-fold configuration was performed for all the models, in order to partition the dataset into train and test sets. First, classification models were evaluated and compared considering all the attributes, except education, and all the observations without missing values. Then, to emphasize the differences of the performance measures while working with an unbalanced and balanced dataset, the classification models were performed again after undersampling the dataset. In this way, the total number of observations decreased to 15016. Successively, the feature selection approach was added to the analysis. The presence of redundant or irrelevant variables was detected through the following approaches: the CFS multivariate filter and the wrapper optimising the F_1 measure.

4.1 Comparing classifiers

The results of the first analysis are shown in table 1.

As we can see, in general all the models have high values of the accuracy measure. The best model is the J48 with an accuracy equal to 85.6%. The worst is the Naive Bayes classifier which has an accuracy equal to

Model	R	P	F-m	A	AUC
J48	0.637	0.748	0.688	0.856	0.884
RF	0.592	0.719	0.649	0.841	0.880
Logistic	0.609	0.735	0.666	0.848	0.904
SVM	0.599	0.734	0.660	0.846	0.764
MLP	0.664	0.655	0.659	0.829	0.874
NB	0.445	0.715	0.548	0.818	0.886

Table 1. Comparing classifiers - unbalanced dataset. R = recall, P = precision, F-m = F_1 measure, A = accuracy, AUC = Area Under Curve.

Model	R	P	F-m	A	AUC
J48	0.841	0.810	0.825	0.822	0.883
RF	0.796	0.805	0.801	0.802	0.884
Logistic	0.842	0.808	0.824	0.821	0.904
SVM	0.858	0.788	0.822	0.814	0.814
MLP	0.729	0.803	0.765	0.775	0.867
NB	0.565	0.868	0.684	0.739	0.886

Table 2. Comparing classifiers - balanced dataset. R = recall, P = precision, F-m = F_1 measure, A = accuracy, AUC = Area Under Curve.

81.8%. However, the accuracy measure treats every class equally important and it may not be suitable for analysing imbalanced dataset. For this reason, it is important to consider the F_1 measure values. They tend to be lower, reaching a value of 68.8% for the J48 model and a value of 54.8% for the Naive Bayes model. The comparison of the classifiers is also represented by the following ROC-curve. The highest value of the AUC is equal to 0.904, achieved by the logistic model. Overall, J48 and logistic models have the best measure performance in terms of AUC and F-measure.

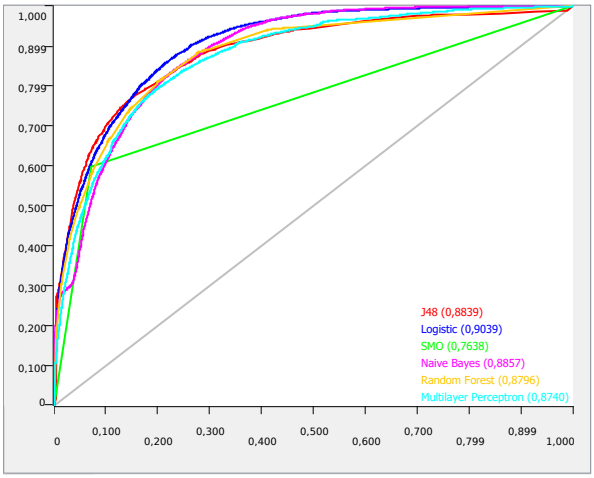

Figure 3. ROC Curve of different classifiers with unbalanced dataset. FPR on x-axis, TPR on y-axis.

4.2 Class Imbalance problem

The performance of classification models changes when taking into account the class imbalance problem.

As table 2 illustrates, the F_1 measure increases for all the models, reaching high values of 82.5% and 82.4% for the J48 and logistic models, respectively. This improvement can be explained by the dataset that now is balanced with respect to the positive (rare) class. There is not a significant difference concerning the AUC values, that are kept at an acceptable level. Also in this case, the best models are J48 and logistic.

4.3 Feature selection

In order to overcome the issues related to the presence of irrelevant attributes, the feature selection approach was performed to the balanced dataset. All the models were performed again, considering CFS multivariate filter and wrapper optimising F_1 measure approaches. Table 3 shows the results of the CFS multivariate filter.

Model	R	P	F-m	A	AUC
J48	0.857	0.799	0.827	0.821	0.895
RF	0.823	0.786	0.804	0.799	0.875
Logistic	0.850	0.793	0.820	0.814	0.893
SVM	0.878	0.748	0.807	0.791	0.791
MLP	0.820	0.802	0.811	0.809	0.892
NB	0.396	0.865	0.543	0.667	0.877

Table 3. Comparing classifiers - balanced dataset & Correlation Feature Selection. R = recall, P = precision, F-m = F_1 measure, A = accuracy, AUC = Area Under Curve.

Again, J48 appears to be the leading model with the highest value of the AUC and the F_1 measure, that increased to 82.7%. Good performances are achieved also by the logistic model, with slightly lower F_1 measure and AUC, compared to the J48 classifier.

Successively, the wrapper feature selection approach is performed with the optimization of the F_1 measure. Results are shown in table 4.

Model	R	P	F-m	A	AUC
J48	0.845	0.805	0.825	0.820	0.890
RF	0.892	0.776	0.830	0.817	0.899
Logistic	0.841	0.807	0.823	0.820	0.902
SVM	0.828	0.779	0.803	0.797	0.797
MLP	0.856	0.768	0.810	0.799	0.881
NB	0.856	0.767	0.809	0.798	0.876

Table 4. Comparing classifiers - balanced dataset & Wrapper. R = recall, P = precision, F-m = F1 measure, A = accuracy, AUC = Area Under Curve.

Random Forest appears to be the leading model in term of the F_1 measure, keeping AUC at a good level. The values achieved are 83% and 0.902, respectively. It is interesting to emphasize that with this feature selection approach, also the Naive Bayes classifier achieved good performance values in terms of F_1 measure and recall, while in the other settings these measures were much lower.

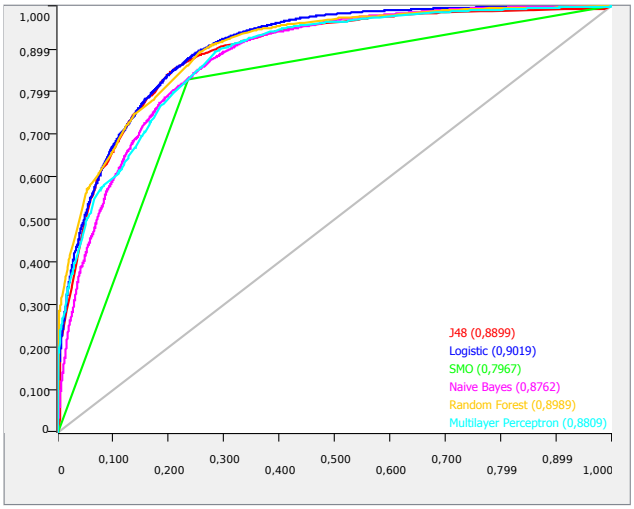

Figure 4. ROC Curve of different classifiers with balanced dataset and Wrapper. FPR on x-axis, TPR on y-axis.

In this case, the logistic model has the highest value of the AUC. Nevertheless, Random Forest appears to be the leading model with high values of the F_1 measure and AUC.

5. Conclusions

The objective of this project consisted in finding a classification model able to predict the income based on some easily accessible socio-economic attributes. The problem was studied in the context of a binary classification task and handled a class imbalance case, which was solved with the undersampling procedure. Finally, feature selection was performed, allowing to find the most relevant attributes.

The models evaluated are: J48, Random Forest, Logistic Regression, Multi-Layer Perceptron, Naive Bayes and Support Vector Machine. Among all the classifiers, the best seemed to be the ones belonging to the heuristic category, which are J48 and Random Forest, and the logistic one. In particular, for almost all of the analyses, J48 was the best model, followed by the logistic regression. Only in the wrapper measures the best model was Random Forest, followed by the logistic one.

Dealing with the class imbalance problem, sampling based approach proved itself effective. In facts, through the undersampling procedure, it was possible to improve significantly the performances of the models, in terms of F-measures.

Through feature selection, instead, the performances of the classifiers did not change significantly. However, CFS multivariate filter was able to decrease the number of relevant input attributes from 14 to 6. In particular, this approach selects the following features:

1. *age*
2. *education.num*
3. *marital.status*
4. *relationship*
5. *capital.gain*
6. *capital.loss*

The wrapper method maximising the F-measure was also applied. The number of attributes selected ranged from 4 to 9, choosing for every model different types of features; only the attribute 'Relationship' appeared in every selection. However, the results are not as valuable as those found with the previous approach. The computational costs were much higher compared to the CFS filter and the performance measures were similar.

The classification analysis provides interesting results that could be explored more in depth in order to

gain reliable and relevant knowledge for different purposes.

References

[1] Adult census income. https://www.kaggle.com/uciml/adult-census-income.

[2] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

[3] Ten techniques to deal with imbalanced classes in machine learning. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/.

[4] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997. Relevance.