Supervised and Non-supervised Learning Algorithm

A comparison between different methods

Dao Ton Son

Contents

1 Abstract 2

2 Dataset 2

3 Data preprocessing 4

3.1 Data imputation 4

3.2 Undersampling 5

3.3 Standardization 5

4 Supervised Learning and Unsupervised learning 6

4.1 Decision Tree 6

4.2 Random Fores and Baggingt 7

5 Unsupervised Learning 8

6 Result 10

6.1 Supervised Learning 10

6.2 Unsupervised Learning 11

7 Discussion 11

8 Conclusion 12

1 Abstract

This project aims to classify individuals' income level, specifically whether they make more than 50 thousand per month. Various supervised machine learning methods, including random forest and decision tree, along with bootstrapping, are employed to develop predictive models. Additionally, unsupervised learning using hierarchical clustering is applied for comparison. The project addresses common data processing challenges such as class imbalance, heavy-tailed distribution, and missing data using oversampling,standardization and mice forest imputation method respecitively. By utilizing different techniques, the project aims to provide insights into the factors influencing income levels and contribute to effective income prediction models. In the end we are able to show the effectiveness of these classic binary classification method in dealing with these type of data.

2 Dataset

The dataset is from the UCI Machine Learning Repository which is the University of California's collection of databases, domain theories, and data producers. Specifically, I use the dataset of Barry Becker from the 1994 Census database. In total, there are more than 48000 observations with 14 features, 6 numerical and 8 categorical features. The final column show the dependent variable which is whether an individual makes more or less than 50 thousand dollar a year. Overall the dataset is high-dimensional with many features can interact with each other such as education vs capital-gain in which the level of education can affect one's ability to invest and make a profit.

Feature	Description	Data Type
age	Individual's age	continuous
workclass	Type of employment	categorical
fnlwgt	Represent the demographic characteristics	continuous
education	Level of education	categorical
education num	Duration of education	continuous
marital status	The current legal marital status of the individual	categorical
occupation	Type of job	categorical
relationship	An individual's relationship to a household or family reference person	categorical
race	Race of an individual	categorical
sex	Individual's sex	categorical
capital gain	Investment's profit	continuous
capital loss	Investment's loss	continuous
hours per week	Work duration of an individual each week	continuous
native country	Where an individual is born	categorical

Looking at the figures below, there is an obvious data imbalances between classes in the same features such as in the case of male and female in sex where number male is dominant over female. One or two class will always dominant over other class. Especially in the case of capital gain and capital loss where many individuals does not invest and the one who does will gain very little or very profitable, however, there is nothing in the middle. This indicates heavy tails distribution and outliers in the data set which will impact negatively on the statistical models. In addition, the out of 48000 observation, around 4 thousands observations have missing value(s). These problem will be addressed in this project.

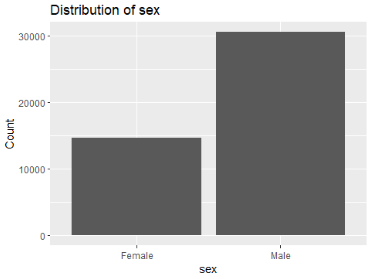


Figure 1

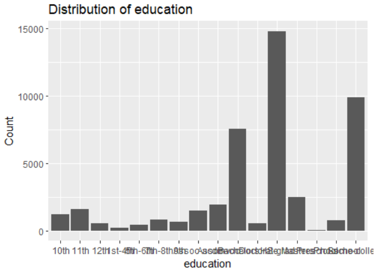


Figure 2

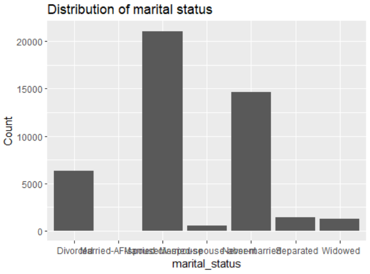


Figure 3

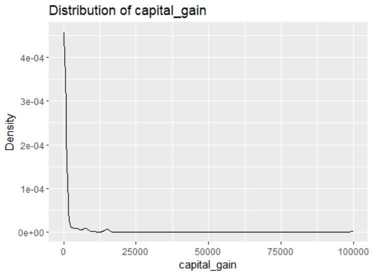


Figure 4

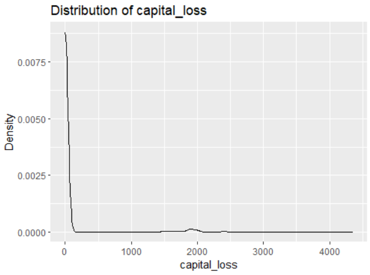


Figure 5

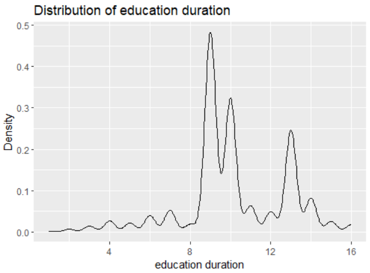


Figure 6

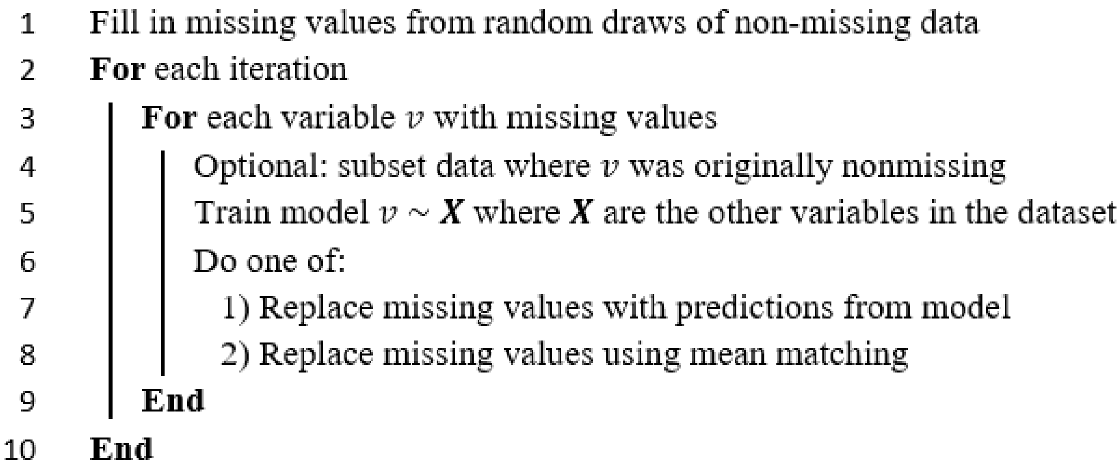
3 Data preprocessing

3.1 Data imputation

The data set contains some problem with almost 4000 missing values, heavy tails distribution and class imbalances in many features. To handles these problem, I apply MICE(Multiple Imputation By Chained Equation) forest and oversampling, respectively.

MICE forest is a combination between MICE and random forest. The method can accommodate nonlinearities and interaction in the dataset and generate efficient, less bias and narrower confidence interval than some of the common imputation method [1] . The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate random forest model. The procedure is as follow.

After this, I achieve two different dataset, one with imputed data and one with non-imputed data where I remove any observations with missing data.



```
Fill in missing values from random draws of non-missing data
For each iteration
    For each variable $v$ with missing values
        Optional: subset data where $v$ was originally nonmissing
        Train model $v \sim \boldsymbol{X}$ where $\boldsymbol{X}$ are the other variables in the dataset
        Do one of:
          1) Replace missing values with predictions from model
          2) Replace missing values using mean matching
    End
End
```

Figure 7: MICE forest algorithm

3.2 Undersampling

When exploring the data, class imbalance, where classes in in the target features are not presented equally (e.g the percentage of male and female in a data set), appears in some of the features. This problem can cause predictive model to be bias toward a class and make the prediction accuracy for minority class worsen. In addition, other problem such as false positve/negative, overfitting and misinterpretation of importance might arise when working with imbalanced data.

To overcome the problems emerge with class imbalance, I apply random undersampling where the minority class is duplicated in order to increase the percentage of the class and the majority class is reduced. I apply the method to variable sex.

3.3 Standardization

Another problem with the adult dataset is that many continuous variable lies in the two extreme ends such as capitalgain or capitalloss. This indicates heavy tails distribution that can lead to disproportionate influence on the statistical measures and modeling techniques, leading to biased results. To address this problem, I apply standardization technique to normalize variables. Standardization or Z-normalization is the process of normalizing every values of the dataset such that the mean of all of the values is 0 and the standard deviation is 1. The new value is calculated using the following formula.

$$z = \frac{x - \mu}{\sigma}$$

where: z is the standardized value, x is the original value, μ is the mean of the data, and σ is the standard deviation of the data.

4 Supervised Learning and Unsupervised learning

In this section, with normalized data, I apply decision tree, random forest and bagging as supervised method to build the predictive model for the adult dataset with and without mice forest imputation, and model with class balance. And for unsupervised learning, I build model using hierarchical clustering to help with the classfication task. I will build the model based on normalized data and normalized data that is undersampled.

4.1 Decision Tree

A decision tree is a flowchart-like algorithm used for classification and regression tasks. It recursively splits the data based on features to make decisions. Decision trees are interpretable and can handle both categorical and numerical data. They are prone to overfitting but can be controlled through techniques like pruning and ensemble methods. Decision trees are widely used for their simplicity and transparency in decision-making. There are two type of decision trees, one is regression and the second one is classification tree. For this problem at hand, I am using the latter. To measure classitication decision tree's impurity, Gini index is implemented. Gini index quantifies the probability of misclassifying a randomly selected element from the node if it were labeled randomly according to the distribution of classes in that node. the method is computationally efficient and tends to favor majority classes, which make it susceptible for imbalanced datasets, hence, it is important to oversampling the the data before building the model. The Gini index is particularly useful when the goal is to build a simple and interpretable decision tree. The trees'performance for binary classfication task are evaluated based on the following performance matrix.

(1)**Accuracy**

$$\frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

4.2 Random Fores and Baggingt

As I mention in the previous section, decision tree can suffer from overfitting, hence, other methods are developed to cope with the problem. One of the method is random forest where we take the average of many decision trees to reduce overfitting and improve the overall predictive performance. Random forest builds an ensemble of decision trees by randomly selecting subsets of the training data and features. Each tree in the forest is trained independently, and the final prediction is obtained by averaging the predictions of all the individual trees. In addition, random forest includes a mechanism to control overfitting by limiting the depth of individual decision trees. This approach helps to mitigate the high variance of individual decision trees and provides better generalization to unseen data. Moreover, random forest has the ability to handle high-dimensional data, capture non-linear relationships, and identify important features for classification tasks.[2]

Another method is bagging or Bootstrap aggregating, instead of using only a subset of features, uses all features to create decision trees. When a split happens, the algorithm randomly selecting a subset of predictors at each split aiming to reduce the correlation between the trees and improve the overall predictive performance. This random feature selection is one of the key differences between bagging and a standard decision tree, where all predictors are typically considered at each split.

Random forest's and bagging's impurity are also measured with Gini index and evaluated with **Accuracy**

5 Unsupervised Learning

Hierarchical clustering is an alternative from k-means clustering that does not necessary need the number of clusters(k) before hand. Hierarchical clustering builds a dendrogram, or tree-like structure, by starting from individual observations (leaves) and gradually merging them into clusters. The algorithm utilizes distance measures to determine the similarity between clusters and combines them based on their proximity. Hence, it is necessarily for the dataset to be normalized so that the distance between data points are on the same scale. For our data set, the continuous variables are standardized with mean zero and standard deviation equal to 1.

The distance between observations is computed using Gower distance. The Gower distance is calculated as the average of partial dissimilarities across instance. We use the folowing equation to measure Gower index.

$$d_{ij} = \frac{\sum_{k=1}^{p} \cdot S_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

where:

$S_{ijk} = 1$ when it is possible to assess the similarity between two observations, it is equal to 0 otherwise.

For qualitative variables, $S_{ijk} = 1$ when the observations are in the same class.

For quantitative variables,

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{K_k}$$

.

where K_k is the variation change of the variable X_k The Gower distances has range from 0 to 1. $d_{ij} = 1$, When all the observations belong to the same class and have identical quantitative values, indicating a **perfect similarity**.

$d_{ij} = 0$ When no pair of observations shares the same class and, for each quantitative variable, the observations have opposite values, indicating maximum diversity**maximum diversity**

In addition to calculating Gower distance, I also select the linkage method before implementing the clustering algorithm which is important because different linkage methods can yield different clustering results. The choice of linkage method impacts how the distances between clusters are calculated, which in turn affects how clusters are merged during the hierarchical clustering process. For the adult dataset, I implement average linkage method. The latter do the same but record the average of the results. I implement hierarchical clustering on only normalized dataset and normalized data set that goes through undersampling. Since gower distance in Rstudio can only handle to a certain amount of data, I randomly sample a subset of around 2300 samples from each data set mentioned above.

6 Result

6.1 Supervised Learning

Accuracy for each model is recorded and shown in Table 1

Table 1: Accuracy for each model

	Decision Tree	Randon Forest	only undersampling
No imputation/undersampling	0,837	0,856	0,861
Only imputation	0,849	0,864	0,869
Only undersampling	0,886	0,885	0,871

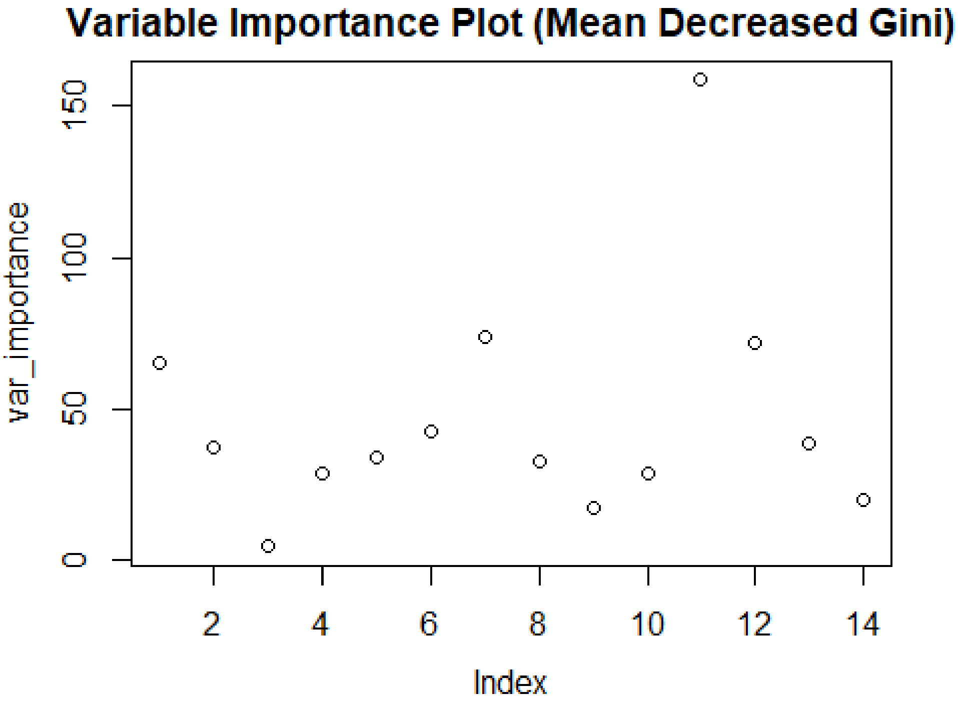


Figure 8: Feature importance of random forest with undersampling data

6.2 Unsupervised Learning

Figures 9, 10, 11, 12 show the results for unsupervised method.

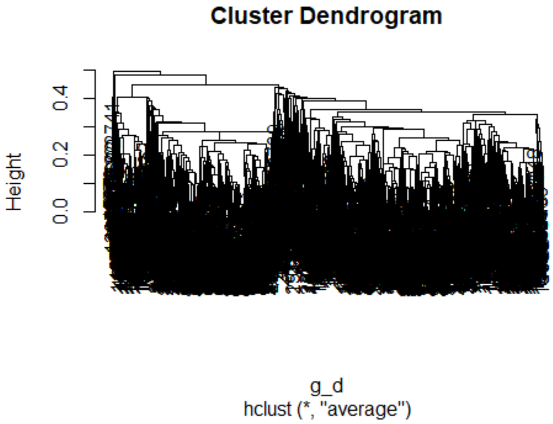


Figure 9: Dendrogram of undersampling data

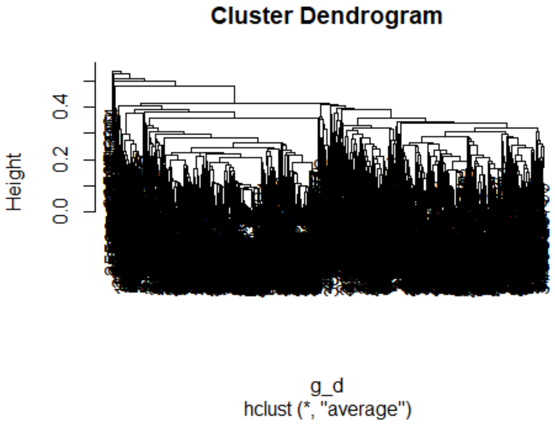


Figure 10: Dendrogram of imbalance data

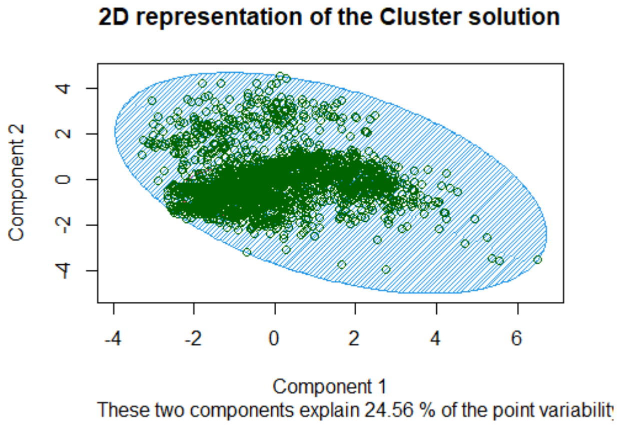


Figure 11: Hierarchical clustering for balance data

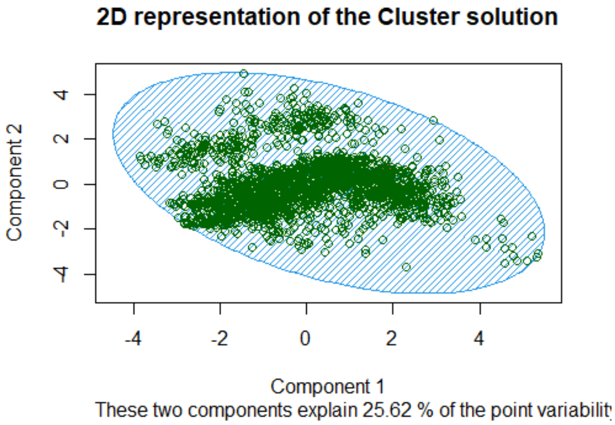


Figure 12: Hierarchical clustering for imbalance data

7 Discussion

In supervised learning model, overall, the performance between models are similar (around 0.86) with the exception of decision tree, showing evidence that overfitting took place in decision tree. With the help of imputation or undersampling, the model generate less bias and error than the model without imputation nor undersampling. Throughout the whole 3 model, undersampling improved accuracy the most with the best result is in random forest model (0.885). By looking at figure 8, it suggests that capital-gain responsible the most for the prediction of this model following by age, occupation and capital-loss.

For unsupervised model, the gower distance shows that it is susceptible to the susceptible to the characteristics of adult dataset which is tails heavy and imbalanceness. It is difficult to notice any cluster in the dendrogram of both model in figure 9, 10. In addition, 2D representation of cluster solution cant separate the clusters of points in a meaningful way.

8 Conclusion

Overall, for the adult dataset of Barry Becker from the 1994 Census database, the supervised models clearly out perform unsupervised model in binary classification. Even though the dataset is characterised with heavy tails distribution as well as class imbalance in many features, classic binary classification methods are still able to cope with the problem. Further more, data imputation and undersampling imbalance data do contribute to cope with those problems. By leveraging their strengths in handling these issues, these models have demonstrated good performance and can be considered reliable solutions for classification tasks in such scenarios.

References

[1] Anoop D Shah et al. “Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study”. In: *American journal of epidemiology* 179.6 (2014), pp. 764–774.

[2] Louis Capitaine, Robin Genuer, and Rodolphe Thiébaut. *Random forests for high-dimensional longitudinal data*. 2019. arXiv: `1901.11279 [stat.ME]`.