ML Project 2: Probabilistic Supervised Classification

Elisa Marson

November 22, 2022

1 Introduction

This project is going to analyse and classify a dataset about employee turn-over.

1.1 Problem Description

Employee turn-over (also known as "employee churn") is a costly problem for companies. The true cost of replacing an employee can often be quite large. A study by the Center for American Progress found that companies typically pay about one-fifth of an employee's salary to replace that employee, and the cost can significantly increase if executives or highest-paid employees are to be replaced.
In other words, the cost of replacing employees for most employers remains significant. This is due to the amount of time spent to interview and find a replacement, sign-on bonuses, and the loss of productivity for several months while the new employee gets accustomed to the new role.

1.2 Methodology

Given that we have data on former employees, this is a supervised classification problem where the label is a binary variable, where 0 represent an active employee and 1 a former employee. The dataset we are going to explore is Employee Analysis — Attrition Report, which contains 1470 observation and 35 variables.
We are going to use different classification algorithms (Logistic Regression, Linear Discriminant Analysis, Naive Bayes algorithms, Boosting and Bagging) first using all the variables, then we are going to apply different methods to perform feature subset selection (univariate filter feature subset selection, multivariate filter feature subset selection and multivariate wrapper feature subset selection).

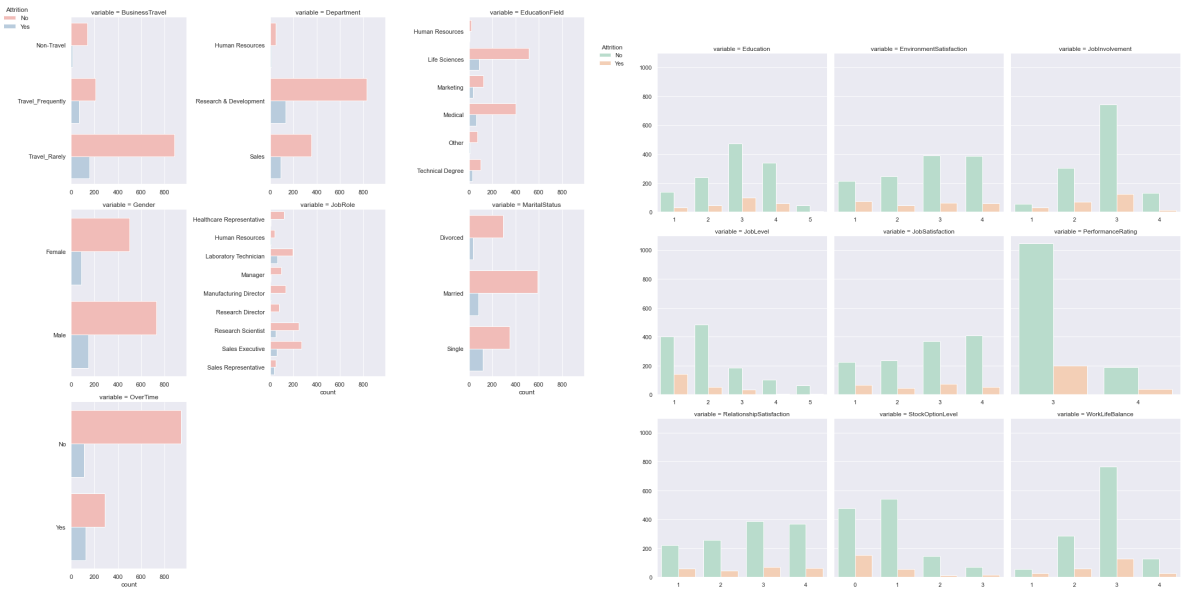
2 Data

First of all, I am gonna explore the variables of the dataset, to see if all of them are useful.

- The variable *Attrition* is our target variable.
- The variable *EmployeeNumber* is dropped since it is just an identifier.
- Based on the description of the dataset from the creator, the following variables are ordinal: *Education*, *EnvironmentSatisfaction*, *JobInvolvement*, *JobSatisfaction*, *PerformanceRating*, *RelationshipSatisfaction*, *WorkLifeBalance*.
- After checking the dataset, *JobLevel* and *StockOptionLevel* should also be ordinal.
- After checking for variables with constant value, the following were removed: *EmployeeCount*, *Over18* and *StandardHours*.

2.1 Data Visualization

Now let's explore all the variables I am going to use for this classification problem through some visualization. The variables are going to be plot grouped by their type: nominal variables, ordinal variables and numerical variables.



(a) Nominal Variables

(b) Ordinal Variables

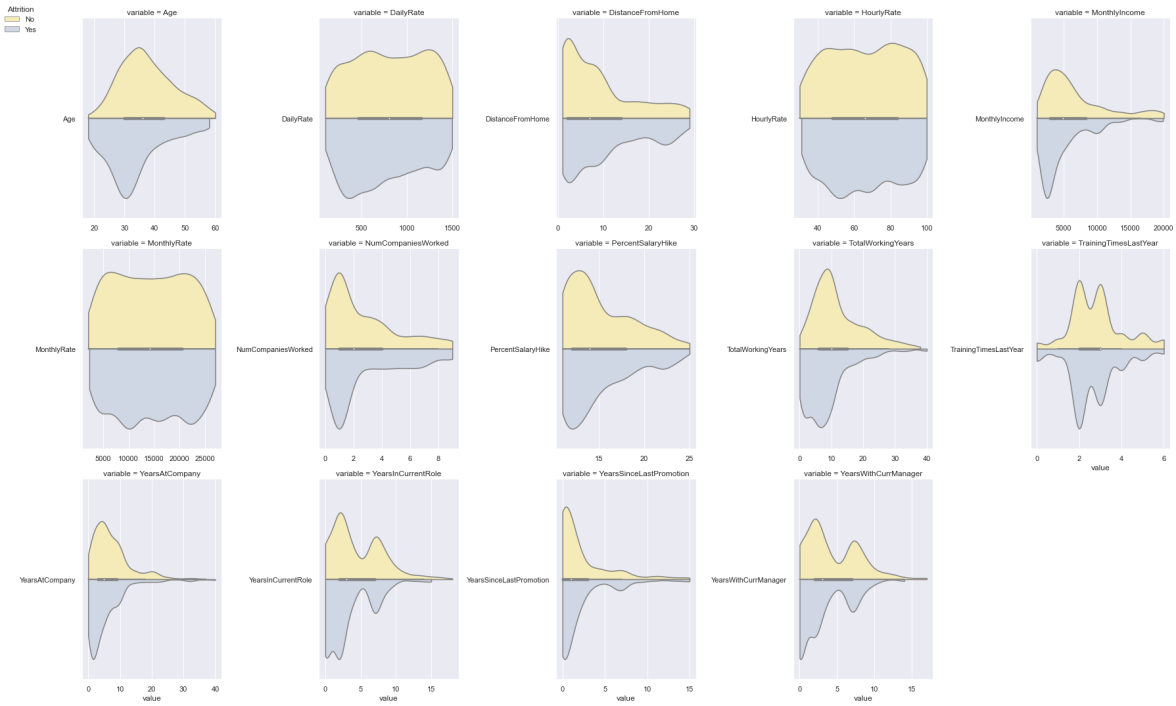


Figure 2: Numerical Variables

Through visualization we can already understand a few characteristics about our data and, consequently, about the employees. Most of the them:

- Are male
- Rarely have business travel
- Are in Research and Development department
- Are from Life Sciences field followed by Medical field

- Do not overtime
- Huge number of employees have high 3-JobInvolvement and better 3-WorkLifeBalance
- Average age of employees is 36 and average monthly income is 6503
- Half of the employees live less than 7 kms from the office

Moreover, we can see that for the variable *PerformanceRating* there are only two levels (3-Excellent and 4-Outstanding), something that we have to keep in mind if this variable will result to be important for our predictions. Let's now focus on our target variable. *Attrition* happens more:

- to single employees
- to laboratory technician, sales executive and research scientist
- to those with with business travel rarely

Others characteristics about employees and attrition:

- Most employees who attrite are on early 30s while most who stay are on mid 30s
- Based on the density plot of the numeric variables, distribution of employees who attrite and not, are almost the same except for *Age*, *MonthlyIncome*, *TotalWorkingYears*, *YearsAtCompany*, *YearInCurrentRole* and *YearsWithCurrManager*.
- Huge number of employees who leave have around 2500 monthly rate.
- The length of stay in the company is more or less three 2.5 years

2.2 Data Preparation

First of all, there are categorical variables that are binary, *Gender*, *OverTime*, *PerformanceRating* and the target one *Attrition*, so I mapped to 0 and 1. As said before, *Performance Rating* has only two values, 3 & 4. I modified the values such that 0 replaces 3 and 1 replaces 4.

Regarding the nominal variables, the followings need to be dummy encoded: *BusinessTravel*, *Department*, *EducationField*, *JobRole* and *MaritalStatus*.

Lastly, ordinal variables do not need to be dummy encoded since we would lose the ordered nature of the variables which is important in this case. I assumed that the differences in levels are more or less similar and that we can treat them as continuous.

The data have also been scaled and, after checking collinearity, I dropped the following variables: *MonthlyIncome*, *Department_Sales* and *JobRole_Human Resources*. Before splitting into train and test, there is one last thing I had to take care of.

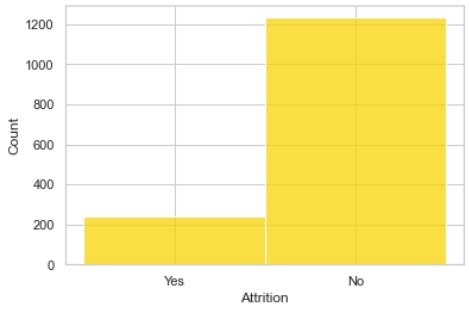


Figure 3: Distribution of Attrition

As we can clearly notice form the graph above, this is an imbalanced dataset. To avoid the consequences this problem can raise in the predictions, I used the SMOTE function to oversample the minority class. Now we are ready to proceed with the implementation of the classification algorithms.

3 Results and Discussion

As stated in the begging, I used 4 different kind of algorithms plus 4 types of Naive Bayes algorithms; each one of them has been used in 4 different analysis.

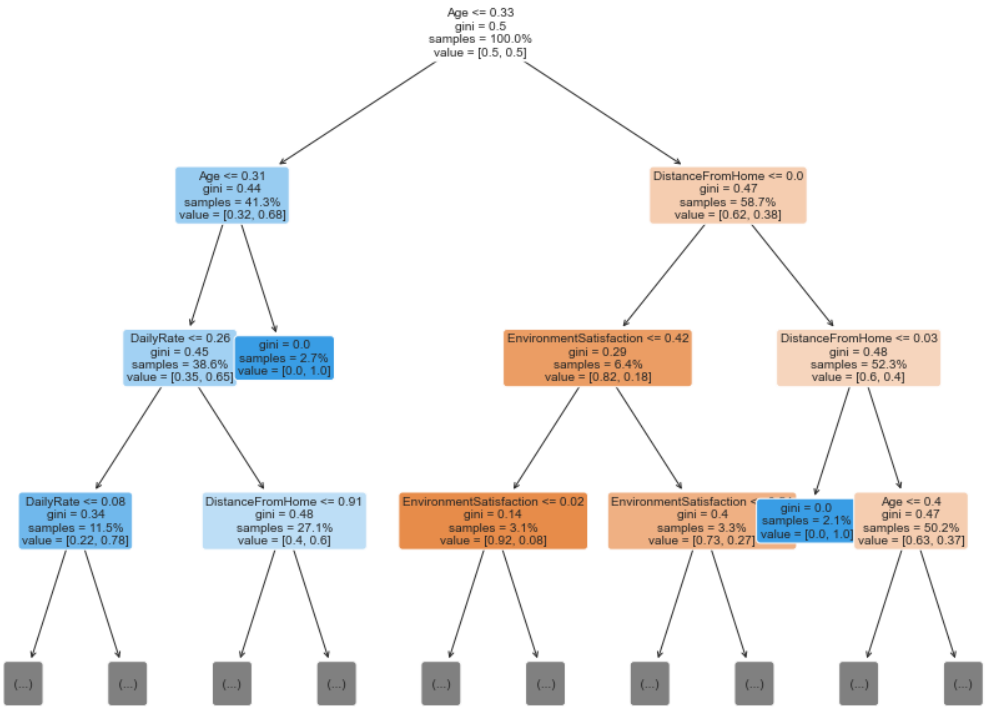
In the first analysis I used all the variables; in the second one I selected a subset of variables using the Chi Squared test; in the third one I selected the subset based on the Mutual Information score of the variables; in the last one I used the Sequential Forward Selection method.

The results are listed in the table below.

	All Variables	ChiSquared	MI Score	SFS Method
Logistic Regression	0.768707	0.671202	0.732426	0.718821
Linear Disciriminant Analysis	0.759637	0.671202	0.723356	0.712018
Naive Bayes	0.560091	0.798186	0.551020	0.632653
Multinomial NB	0.673469	0.691610	0.648526	0.662132
Complement NB	0.673469	0.691610	0.648526	0.641723
Bernouilli NB	0.721088	0.730159	0.705215	0.526077
AdaBoost	0.873016	0.723356	0.843537	0.682540
Bagging	0.879819	0.759637	0.854875	0.748299

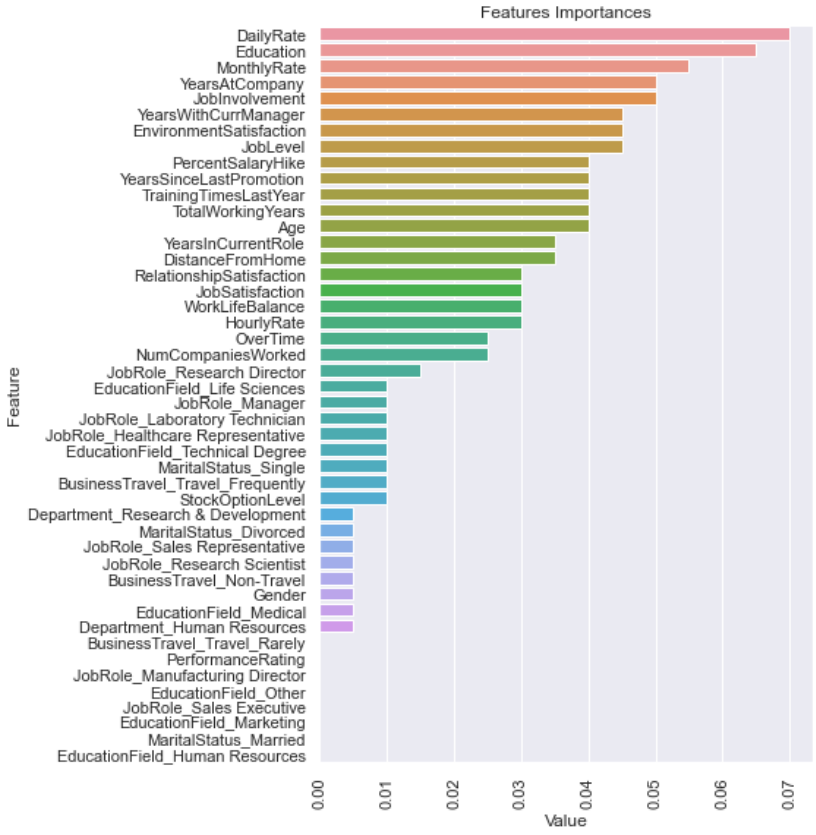
Figure 4: Accuracies Table

Looking just at the accuracy scores of the algorithms, the one that gives us the highest accuracy is the Bagging classifier used with all the features; let's take a look at one tree built by the algorithm (only the first part of the tree is shown).



The first feature we find in the first decision node is *Age*, and in the following ones we find the variables *DistanceFromHome*, *DailyRate* and *EnvironmentSatisfaction*. It seems so that to predict if an employee is going to quit the job or not it is important to look at his/her entire satisfaction both in the job life and the private one; it doesn't only depend on the characteristics of the work space or the job role for example. These are of course important but they have to be sustained by a comfortable and stable private life; indeed a young employee is more likely to quit.

The other classifier that has a high accuracy is the AdaBoost; let's check which are the most important features used by this algorithm.

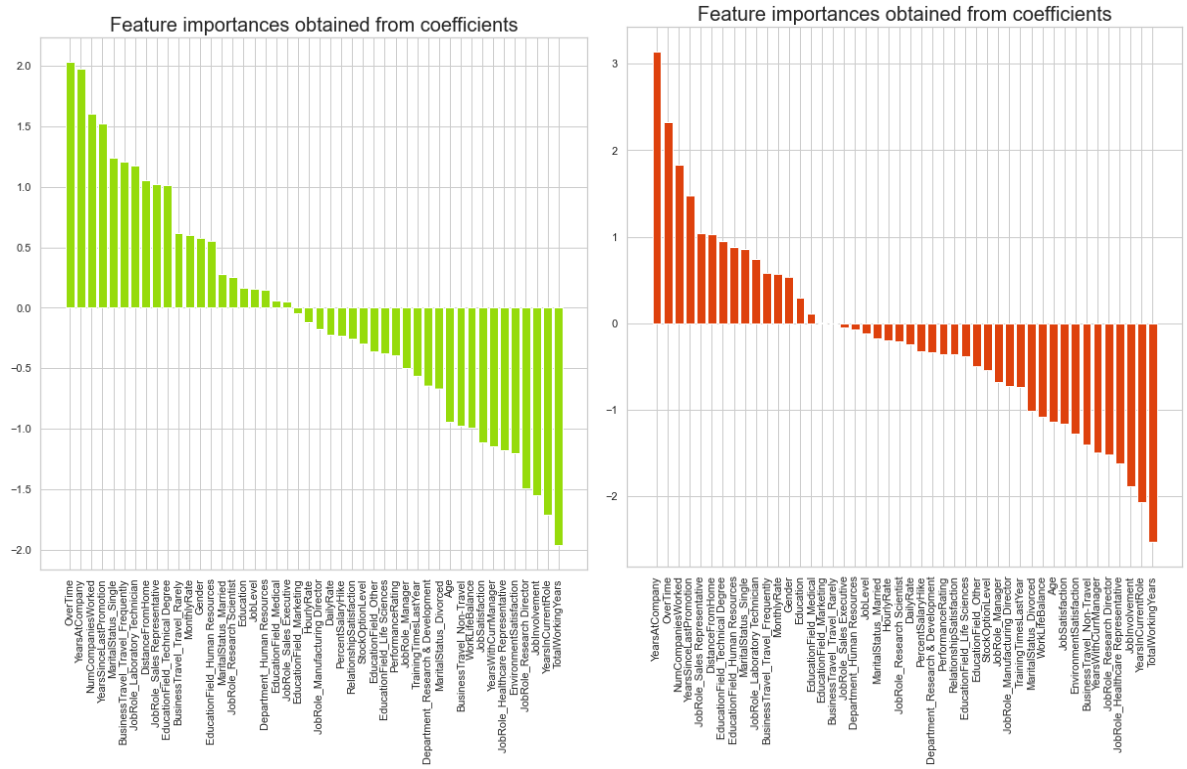


DailyRate becomes now the most important feature; not so surprising that salary has a heavy weight on the decision of an employee; indeed the third most important variable is *MonthlyRate*.

A part from *Education*, that we find in the second place, in this case the AdaBoost classifier takes into account variables related more to the job life of the employee; in order of importance, after the ones we already discuss about, we find *YearsAtCompany*, *JobInvolvement*, *YearsWithCUrrentManager* and, as in bagging, is important also the *EnvironmentSatisfaction*.

The variable *Age*, that before was so important, is now only in the 13th place. Notice also that the variable *OverTime* is only in the 20th place; this observation is going to be useflu later.

Let's now go to explore the results from the others algorithms implemented. Even though their accuracy was worse, we can still obtain useful insights. Below are plotted some graphs about features importance of the Linear Regression and and the Linear Discriminant Analysis.



(a) Logistic Regression Feature Importance

(b) Linear Discriminant Analysis Feature Importance

From the graphs above, we notice that, in both the classifiers, the variable *OverTime* becomes highly important to predict if an employee is going to quit, together with *YearsAtCompany*.

On the other end, both the classifiers give high importance to the variables *YearsInCurrentRole* and *TotalWorkingYears* to predict that an employee is going to remain at the company. These variables together with *JobInvolvement* seems to say that the higher the importance of the employee in the company, the lower the chances he/she is going to quit. According with this line of thought, the variable *YearsSinceLastPromotion* is important to predict the quitting of an employee: if the company doesn't take into account his/her effort in the job, he/she is probably going to search for another place where the efforts are recognize; or maybe this kind of job was not in his/her skills and a lack of promotion is a symptom of not suited employee for the job.

Lastly is curios how the variable *MaritialStatus_Single* is useful for the Logistic Regression classifier to predict that an employee is going to quit: also according to this algorithm, personal and job life are very linked together.

Finally, let's move on analysing the Bayes algorithm. The one that performs best according to accuracy is the Naive Bayes with the subset of features selected with the ChiSquared test: it reaches the accuracy of 79.8%, being the third best.

The 5 features considered in this case are listed below.

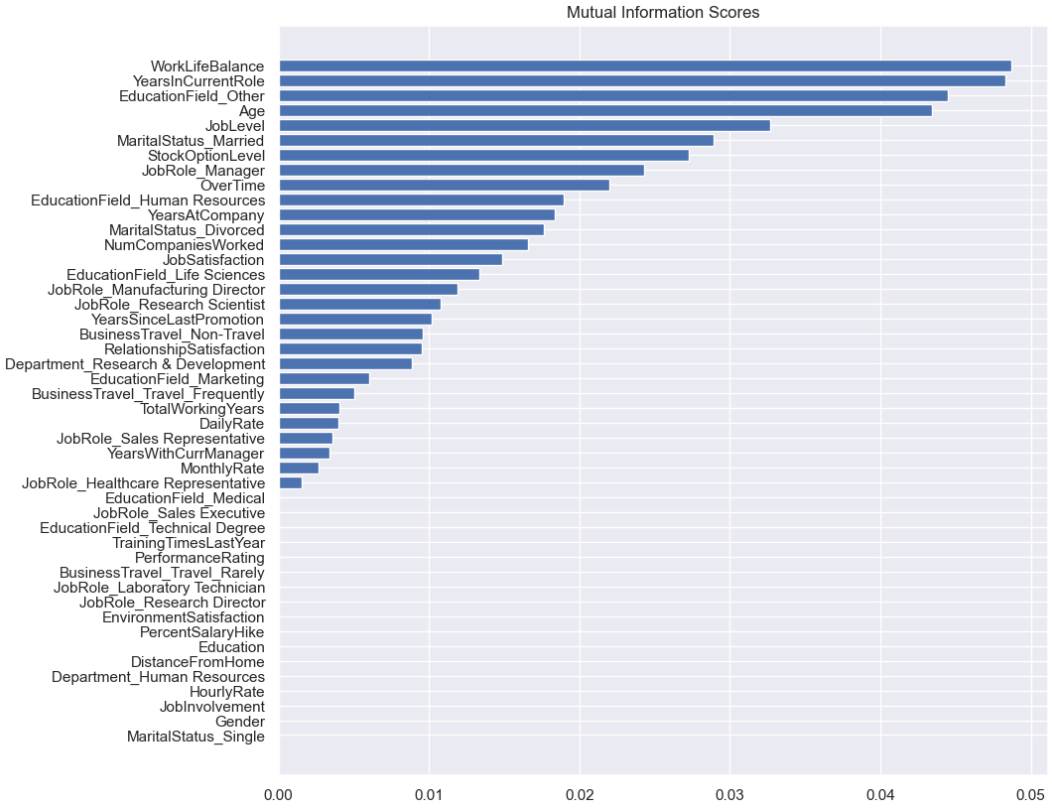
	Specs	Score
12	OverTime	63.845067
42	JobRole_Sales Representative	34.290268
45	MaritalStatus_Single	30.771669
25	BusinessTravel_Travel_Frequently	15.816623
8	JobLevel	12.094895

Another time we find here that *OverTime* is the most important variable of all and again we encounter *MaritialStatus_Single*.

The importance of the variable *OverTime* in the last 3 classifiers analyzed actually can be connect to the line of thought we were following while analyzing the Bagging classifier.

If an employee has more time to dedicate to his/her private life in order to build a more comfortable and stable environment, is more likely to not quit. In the opposite case, doing overtime means having less time to do this and so being more unsatisfied with his/her life in general: quitting a job that is draining your private life could be the first step to a better life.

Lastly, let's check which variables have been chosen with the Mutual Information score method, even though this method didn't give us outstanding performances.



Not so surprisingly, the variable *WorkLifeBalance*, that can be seen as a summary of the variables between work and private life, is in this case the most important, supporting the line of thought followed until here.

4 Conclusions

The Bagging classifier is the one that performs better and seems to take into account the right variables about building a balanced life between the job one and personal one.

Anyway is followed closely by the AdaBoost, that on the other hand focus more on the variables that represent the job life of an employee. Moreover it has a better balance between precision and recall:

- Precision: 91.9%, Recall: 93.4% – AdaBoost
- Precision: 88.4%, Recall: 98.9% – Bagging

Since Precision identifies the proportion of correctly predicted positive outcome, it means it is more concerned with the positive class than the negative class.

In conclusion, Bagging performs slightly better in terms of accuracy but, if predicting the negative class is also important, the AdaBoost would be preferred.