MACHINE LEARNING II

INDIVIDUAL ASSIGNMENT

HR ANALYTICS

TABLE OF CONTENTS

INTRODUCTION

This case study aims to model the probability of attrition of each employee from the HR Analytics Dataset, available on Kaggle. Its conclusions will allow the management to understand which factors urge the employees to leave the company and which changes should be made to avoid their departure. All the files of this project are saved in a GitHub repository.

The libraries used for this project include: *pandas* and *numpy* for data manipulation, *matplotlib.pyplot* and *seaborn* for plotting, *spicy* for preprocessing, and *scikit-learn* for Machine Learning.

DATA PREPARATION

The dataset is stored in the GitHub repository as a CSV file: *turnover.csv*. The file is loaded directly from the repository.

A. Exploratory Data Analysis

The first stage of this analysis is to describe the dataset, understand the meaning of each variable, detect possible patterns and perform the necessary adjustments to ensure that the data will be proceeded correctly during the Machine Learning process.

1. Dataset Description

The dataset consists in 14,999 rows and 10 columns. Each row represents an employee, and each column contains one employee attribute. None of these attributes contains any *NA*.

Variable	Type	Range	Definition
satisfaction_level	Float	0 to 1	Employee satisfaction level.
last_evaluation	Float	0 to 1	Employee last evaluation score.
number_project	Integer	2 to 7	Number of projects handled by the employee.
average_montly_hours	Integer	96 to 310	Average monthly hours worked by the employee.
time_spend_company	Integer	2 to 10	Number of years spent in the company by the employee.
Work_accident	Boolean	0 or 1	Flag indicating if the employee had a work accident.
Left	Boolean	0 or 1	Flag indicating if the employee has left the company. This is the target variable of the study, the one to be modelled.
promotion_last_5years	Boolean	0 or 1	Flag indicating if the employee has been promoted within the past 5 years.
department	Categorical	10 values	Initially *sales*, renamed as *department*. Department of the employee: *Sales, Accounting, HR, Technical, Support, Management, IT, Product Management, Marketing, R&D.*
salary	Categorical	3 values	Salary level of the employee: *Low, Medium, High*.

Figure 1: Variables of the HR Analytics Dataset

2. Key Findings

The objective of this study is to build a model to predict the value of the variable *left*, based on the other variables available. A first inspection reveals that 23.8% of the employee listed in the dataset have left the company. The dataset is not balanced, which might introduce some bias in the predictive model. The Synthetic Minority Oversampling Technique (SMOTE) has been used at the end of the study to compare the model with another one developed from an over-sampled dataset.

A closer look to the means of the variables allow to highlight the differences between the employees who left the company and those who stayed. Employees who left the company have:

- a lower satisfaction level: 0.44 vs 0.67.
- higher average monthly working hours: 207 vs 199.
- a lower work accident ratio: 0.05 vs 0.18.
- a lower promotion rate in the past 5 years: 0.01 vs 0.03.

The salary level seems to have a great impact on the employee turnover, as higher salaries tend to stay in the company (7% of turnover), whereas lower salaries tend to leave the company (30% of turnover). Departments, even with different turnover rates, don't seem to have a significant impact on the employee departure.

Employees with very low satisfaction level (below 0.12) obviously leave the company. A risky zone is when employees rates their satisfaction just below 0.5 (between 0.36 and 0.46). Employees also tend to leave the company when they become moderately satisfied (between 0.72 and 0.92).

Employees with low evaluation scores tend to leave the company (between 0.45 and 0.57). A large number of good employees (scores higher than 0.77) leave the company, maybe to get a better opportunity. Interestingly, the ones with very low scores seem to stay.

Employees with really low numbers of hours per month (below 125) tend to stay in the company, whereas employees working too many hours (above 275 hours) have a high probability to leave the company. A *safe* range is between 161 and 217 hours, which seems to be ideal to keep employees in the company.

The main observation regarding the number of projects is that employees with only 2 or more than 5 projects have a higher probability to leave the company. It also seems that employees with 3-6 years of services are leaving the company. Employees with a work accident tend to stay in the company. And employees with a promotion within the past 5 years have less propensity to leave the company.

No strong correlation appears in the dataset. However, it is possible to see clear groups when looking at the relationships of pairs of variables: Number of Projects vs Average Monthly Hours, Number of Projects vs Last Evaluation, Last Evaluation vs Average Monthly Hours, Last Evaluation vs Satisfaction.

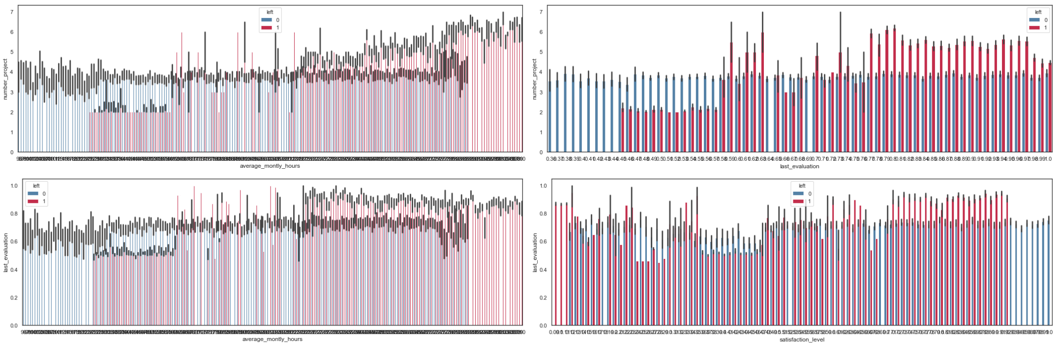


Figure 2: Bar Plots of interesting pairs of variables, highlighting possible groups

B. Encoding, Scaling and Skewness

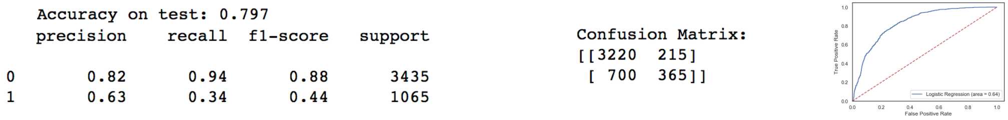
For the model to proceed with the data efficiently, the categorical variables *salary* and *department* have been encoded. As the values of *salary* have an order, they have been encoded into integers within the same variable. For *department*, as the values have no specific order, they have been encoded into individual variables with boolean values. Thus, the dataset has been transformed from 10 variables to 19 variables. Numerical variables *average_monthly_hours*, *last_evaluation* and *satisfaction_level* are scaled between 0 and 1 to remove any influence of their difference in value ranges on the model. They have also been checked for skewness, without a real change on their shape.

C. Train/Test Split

The dataset will be split randomly into Train and Test sets, with ratio 70|30. This method will be used at each step of the feature engineering, before the modelling steps.

BASELINE

A logistic regression algorithm will be used to develop this classification model. The baseline results of the model return an accuracy of 0.797, which is acceptable. However, the results regarding employees who left the company - our main objective - aren't so satisfactory, as they present a very low recall of 0.34, which means that only 34% of the employees who left the company were detected. These results should improve significantly after the Feature Engineering phase for the model to be satisfactory.



```
          Accuracy on test: 0.797
          precision    recall  f1-score   support

0           0.82       0.94      0.88      3435
1           0.63       0.34      0.44      1065
```

```
Confusion Matrix:
[[3220  215]
 [ 700  365]]
```

Figure 3: Baseline Results

FEATURE ENGINEERING

A. Cross Validation Strategy

The model will be cross-validated using a 10-fold cross validation method returning the average accuracy. This method will be applied at every modelling step, to ensure that the model is not biased by the training set split.

B. Feature Construction

In order to improve the model results, a set of features will be created and modified to describe more accurately the characteristics and patterns of the data.

1. Bin Satisfaction Level

Based on the EDA, the Satisfactory Level is binned and one hot encoded in 6 bins: (0.00, 0.11], (0.11, 0.35] , (0.35, 0.46] , (0.46, 0.71] , (0.71, 0.92] , (0.92, 1.00]. The new feature is then one hot encoded. This step increases the accuracy of the model to 0.914. The feature is accepted.

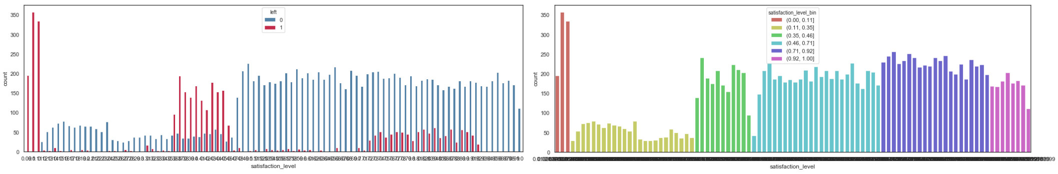


Figure 4: Satisfaction Level Bar Plot, Before and After Binning

2. Bin Last Evaluation

Based on the EDA, the Last Evaluation is binned and one hot encoded in 4 bins: (0.00, 0.44], (0.44, 0.57] , (0.57, 0.76] , (0.76, 1.00]. The new feature is then one hot encoded. This step increases the accuracy of the model to 0.936. The feature is accepted.

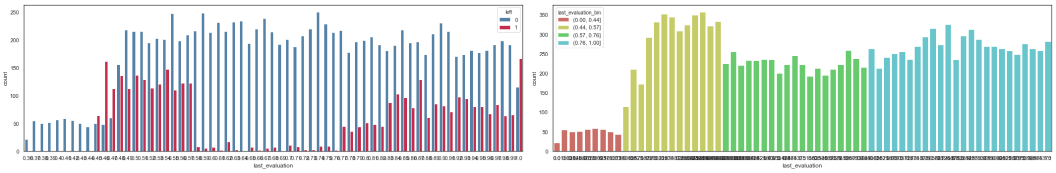


Figure 5: Last Evaluation Bar Plot, Before and After Binning

3. Bin Average Monthly Hours

Based on the EDA, the Average Monthly Hours is binned and one hot encoded in 7 bins: (0, 125], (125, 131] , (131, 161] , (161, 216] , (216, 274] , (274, 287] , (287, 310]. The new feature is then one hot encoded. This step increases the accuracy of the model to 0.945. The feature is accepted.

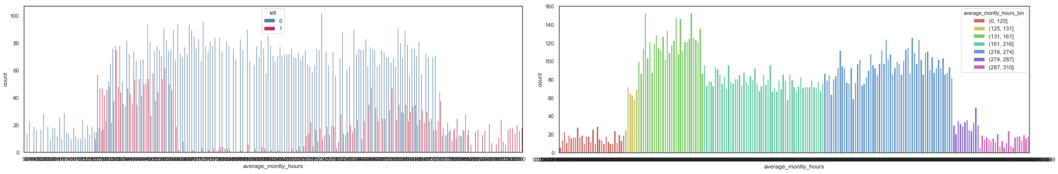


Figure 6: Average Monthly Hours Bar Plot, Before and After Binning

4. Categorize Number of Projects

Based on the EDA, the Number of Projects can be categorized into 4 categories: too low, normal, too high, extreme. The new feature is then one hot encoded. The step increases the accuracy of the model to 0.950. The feature is accepted.

5. Categorize Time Spent in Company

Based on the EDA, the Time Spent in Company can be categorized into 4 categories, related to the rate of departure: no departure, low departure, high departure, very high departure. The new feature is then one hot encoded. The step increases the accuracy of the model to 0.956. The feature is accepted.

6. Cluster by Number of Projects and Average Monthly Hours

Based on the EDA, the employees can be cluster by Workload, based on the Number of Projects and Average Monthly Hours, into 5 categories: very low, low, normal, high, extreme. The new feature is then one hot encoded. The step increases the accuracy of the model to 0.959. The feature is accepted.

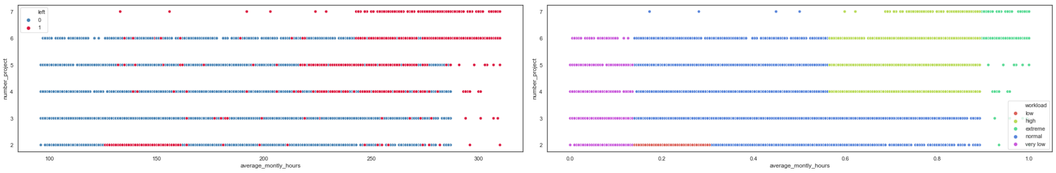


Figure 7: Number of Projects by Average Monthly Hours Scatter Plot, Before and After Clustering

7. Cluster by Number of Projects and Last Evaluation

Based on the EDA, the employees can be cluster by Project Performance, based on the Number of Projects and Last Evaluation, into 4 categories: very low, low, normal, high. The new feature is then one hot encoded. The step decreases the accuracy of the model to 0.958, but the 10-fold cross validation average accuracy increases from 0.958 to 0.960. The feature is kept, even if it is not clearly defined if it has an impact on the model accuracy. The Feature Selection phase might later clarify its importance.

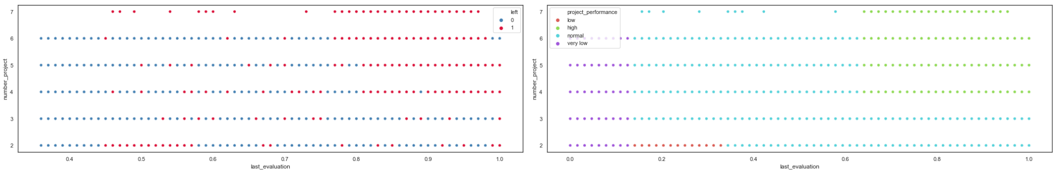


Figure 8: Number of Projects by Last Evaluation Scatter Plot, Before and After Clustering

8. Cluster by Last Evaluation and Average Monthly Hours

Based on the EDA, the employees can be clustered by Efficiency, based on the Last Evaluation and the Average Monthly Hours, into 4 categories: very low, low, normal, high. The new feature is then one hot encoded. The step increases the accuracy of the model to 0.960. The feature is accepted.

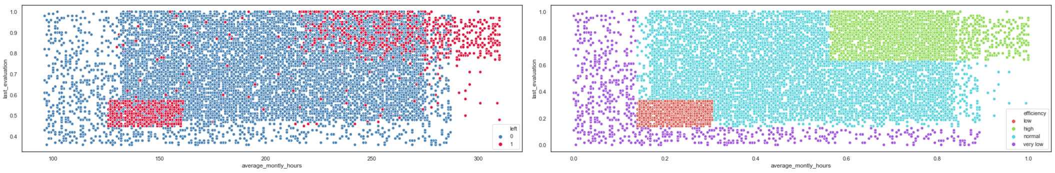


Figure 9: Last Evaluation by Average Monthly Hours Scatter Plot, Before and After Clustering

9. Cluster by Last Evaluation and Satisfaction Level

Based on the EDA, the employees can be clustered by Attitude, based on the Last Evaluation and the Satisfaction Level, into 7 categories: very unhappy, unhappy, low performance, unhappy and low performance, normal, happy and high performance, very happy. The new feature is then one hot encoded. The step increases the accuracy of the model to 0.964. The feature is accepted.

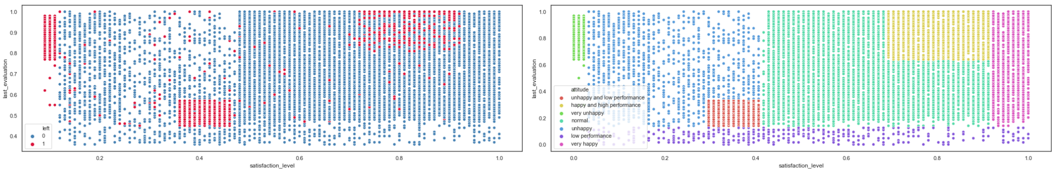


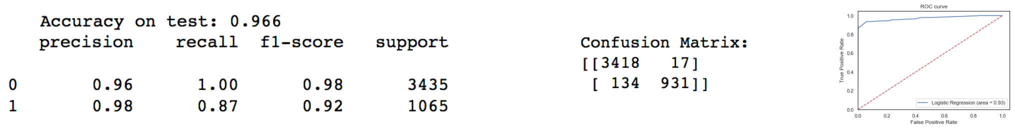
Figure 10: Last Evaluation by Satisfaction Level Scatter Plot, Before and After Clustering

C. Feature Selection

The dataset resulting from the Feature Engineering phase contains 58 features, with a model reaching the accuracy of 0.964. The Feature Selection phase aims to reduce the number of variables used by the model. The Recursive Feature Elimination (RFE) method is used to select the most relevant features for the model. It returns a list of 15 features which should be sufficient to our model.

FINAL METRIC

The final model is tested with the 15 selected features and returns the accuracy of 0.966. The recall for employees who left the company now reaches 87%, which will allow the management to better predict which employees have a high probability to leave.



```
          Accuracy on test: 0.966
          precision    recall  f1-score    support                 Confusion Matrix:
                                                                   [[3418    17]
0           0.96        1.00      0.98       3435                   [ 134   931]]
1           0.98        0.87      0.92       1065
```

Figure 11: Final Model Results

Analyzing the 15 selected features and their coefficients will help understanding the underlying reasons for an employee to want to stay or to leave the company.

Features reducing the probability of departure			Features increasing the probability of departure		
Attitude	Normal	-2.539354	Attitude	Unhappy Low Performance	1.774068
Attitude	Unhappy	-2.014794	Attitude	Very Unhappy	1.669274
Attitude	Very Happy	-2.282289			
Satisfaction	(0.92, 1.00]	-2.298017	Satisfaction	(0.00, 0.11]	2.669274
Workload	Normal	-2.076743	Workload	Extreme	2.179019
Efficiency	Very Low	-4.374203	Efficiency	Low	2.247488
Time Spent	No Departure	-1.945288	Time Spent	Very High Departure	2.473079
			Av. Monthly Hours	(287, 310]	2.179019
			Number Project	Extreme	3.922200
Intercept.		-0.591944			

Figure 12: Selected features and corresponding model coefficients

CONCLUSION

This model will allow the company to calculate the probability of an employee to leave the company and to act on key-factors to avoid departures. The satisfaction of employees and the amount of workload they have to bear seem to be important causes of withdrawals. A particular attention on the work-life balance would be crucial to improve the turnover rate.

LIST OF FIGURES AND TABLES