ABODJI Kondi Kalèd
kondi.kaled.abodji@gmail.com
+212 0619733814

REPORT OF ATTRITION ANALYSIS AND PREDICTION

Introduction

In the organisation of companies, it important to be able to know which employee is about to leave and the others who are staying. In this project, the goal is to setup the best model to do that prediction.

Problem in data science : This a binary classification problem

Objectifs :

- Understant the dataset by exploration
- Visualisation of data
- Analyse of data
- Choose the most revelent columns
- Set up 4 differents model
- Analise accurancy throught metrics
- Choose the best model and Extract

Used technologies :

- Sklearn : Machine Learning library « Models, metrics, normalisation and PCA »
- pandas and numpy for data manipulation
- Matplotlib for data visualisation
- joblib for model extraction

I – Data Exploration & Encoding

The first step is to explore the organisation of data to determine the columns, the general appearence of the data, detect the missing values, the none numeric columns. In summary a general view of the dataset. Then we encode the none numerical columns. Here we encode these columns Index(['Attrition', 'BusinessTravel', 'Department', 'EducationField', 'Gender',

```
'JobRole', 'MaritalStatus', 'Over18', 'OverTime']).
```

Then we handle missing values. In this case there is no missing in the dataset.

II – Features Processing PCA

The goal here is to select the most valuable features in the data. To setup that process, We first do an analysis of the ratio Number of Principal Components and Cumulative Variance. Then we choose the number of PC that maximise at least 95 % of the Cumulative Variance. Here 23 Principal Components. Then we construc the new features. All that processing, should be done after data normalisation.

III - *Models*

Now we have a ready dataset, with well design features. Now we split the dataset into Train data and Test data, we choose 20 % of the data as test set. Then we construct the models by fiting, and evaluting efficiency throught metrics.
We will be using 4 different models

-> Logistic Regression
-> Decision Tree
-> Random forest classification
-> Support Vector Machine

And 3 metrics :

- Accuracy_score
- Confision matrix
- classification report

The results of the setting up are in the jupyter notebook. The following step is the base on the metrics results to choose the most accurate moel for this classification problem.

IV – BEST MODEL

We are not really sure about we analyse to do on the metrics to choose the best model. Teh appropriate combinasion. But We mainly base on the classification report and the confusion matrix which give a great deep overview of the models performance. Our choice is then made on the Logistic Regrestion model that seems to give best balanced results. We then extract the .pkl model using joblib library.

V – CONCLUSION

This project was a classification problem and all our steps had the goal to get the most accurate model to predict attrition. We have analys dataset, manipulate them and do calculation on them through pandas, numpy and PCA in the way to have revelante Features. Then we construct 4 different models, evaluate them and compare their performance, with the help of scikit-learn metrics.