

Table of Contents

[Project Overview 3](#_Toc196614325)

[Business Problem 3](#_Toc196614326)

[Objective 3](#_Toc196614327)

[Data Selection and Research 3](#_Toc196614328)

[PREPROCESSING STEPS 6](#_Toc196614329)

[MODEL LEARNING MODEL DEPLOYMENT 7](#_Toc196614330)

[Training and Evaluation 7](#_Toc196614331)

[Model Selection 8](#_Toc196614332)

[WEB APPLICATION : DASH 10](#_Toc196614333)

[MODEL DEPLOYMENT 10](#_Toc196614334)

[CHALLENGES FACED 12](#_Toc196614335)

[Project links 12](#_Toc196614336)

[GitHub Repository 12](#_Toc196614337)

[Web Application 12](#_Toc196614338)

[CONCLUSION 12](#_Toc196614339)

# Project Overview

Topic : Employee Attrition Prediction (topic 4)

## Business Problem

Employee attrition is a problem faces in multiple organisations and has become a significant challenge within organisations , this has lead to increased hiring costs , decrease in productivity and employee morale. By predicting which employees are likely to leave , the companies are able to take targeted measures to improve retention such as addressing issues such as workplace dissatisfaction amongst other factors.

### Objective

This project aims to predict the likelihood of employee attrition in a company based on factors such as : job satisfaction , performance matrix , tenure and other work-place related features. The model will help the Human Resources (HR) team identify at risk employees and implement the necessary retention strategies as a result.

Approach:

1. Develop a ML model to predict employee attrition using a dataset of employee data.
2. Evaluate multiple methods , models will include : Logistic Regression , Random Forest , XGBoost , to select the best performing model.
3. Build a web application , making use of DASH , to allow for employee inputs to be passed in in order to receive attrition predictions.
4. Deploy model onto Render.

# Data Selection and Research

The dataset used for the project is the “IBM HR Analytics Employee Attrition and Performance” sourced from Kaggle. A synthetic dataset created to simulate real-world HR data for attrition analysis.

Dataset information

Dataset title : IBM HR Analytics Employee Attrition and Performance

Dataset link : [IBM HR Analytics Employee Attrition & Performance](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset)

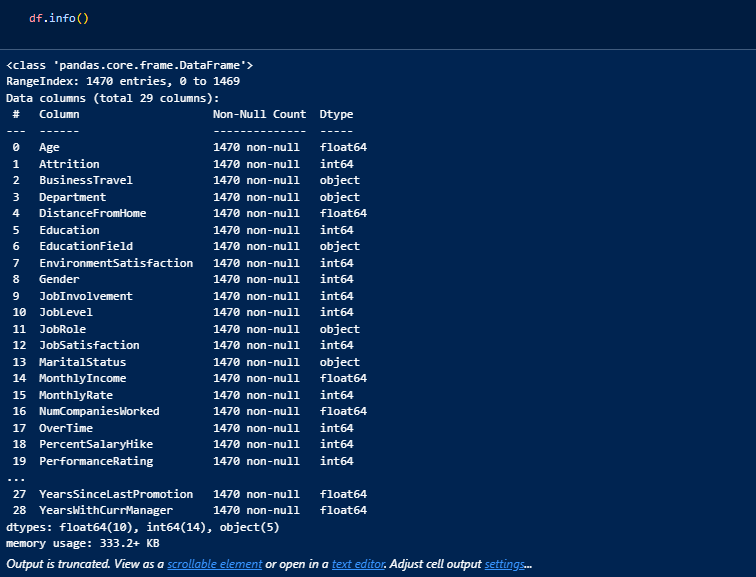
Size : 1470 entries(rows) , 35 columns (Dataset is a manageable size)

Target Variable : Attrition

2. Data Exploration and Preparation

Exploratory Data Analysis (EDA)

As mentioned before the dataset has 1470 rows and 35 columns , with no missing values , seen below



A screenshot of a computer

AI-generated content may be incorrect.

Target Distribution : As seen in the diagram below , Attrition has an imbalance , which will need a balancing technique.



Feature types include both categorical and numerical types , the image below depicts the section of code used to segment these categories. A computer code on a blue background

AI-generated content may be incorrect.

Numerical included

['Age', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager']

Total Numerical Columns: 21

Categorical included

['Business Travel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime']

Total Categorical Columns: 7

The Constant Features

- EmployeeCount: 1

- Over18: Y

- StandardHours: 80

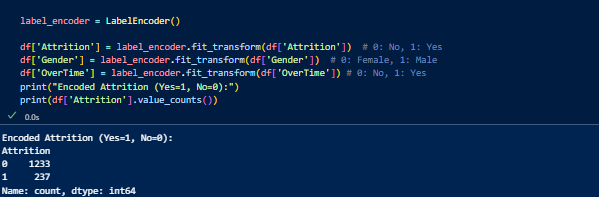
Total Constant Columns: 3

## PREPROCESSING STEPS

1.Dropping Irrelevant Columns



2.Encoding Categorical Variables



3.Scaled Numerical Features

A computer screen shot of a blue background

AI-generated content may be incorrect.

4.Handling Class Imbalance : Made use of SMOTE

A blue screen with white text

AI-generated content may be incorrect.

Scripts were saved under the folder src/prepare\_data.py and src/preprocess\_data.py

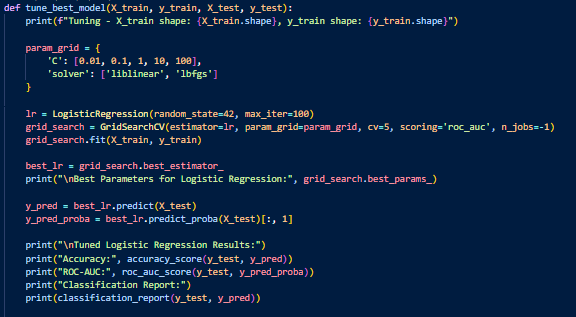
# MODEL LEARNING MODEL DEPLOYMENT

The following models were used for comparison :

* Logistic Regression: a simple and interpretable model used for binary classification
* Random Forest : a tree-based model that handles non-linear relationships.
* XGBoost : a model known for its high performance on imbalanced (its outputs had difficulties displaying)

## Training and Evaluation

GridSearchCV was used for validation during the hyperparameter tuning section



Evaluation Metrics

The measurements used to check the quality of the models predictions are as follows :

* Accuracy
* ROC-AU
* F1-score and recall

A screenshot of a computer

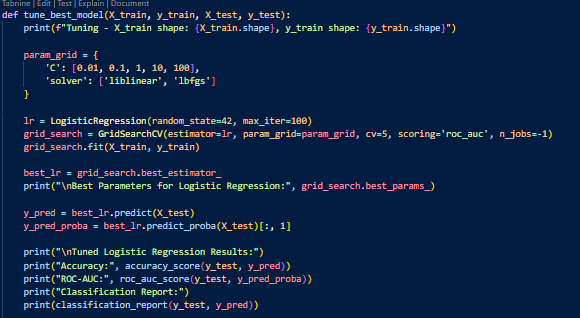
AI-generated content may be incorrect.

### Model Selection

Logistic Regression was chosen based on the following :

* It had the highest ROC-AUC
* Highest recall for class 1 , important for classification of at risk employees
* Best F1-score , balancing the precision and recall for the minority class

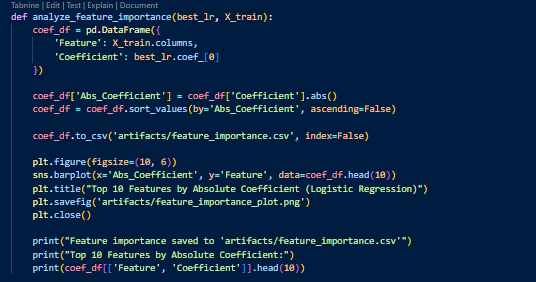
Hyperparameter Tuning

* Tuned the Logistic Regression model using GridSearchCV

Feature Importance

* This show how much a feature contributes to the model’s predictions
* Key : Positive Coefficients = increased attrition

Negative Coefficients = decreased attrition



A screen shot of a computer program

AI-generated content may be incorrect.

# WEB APPLICATION : DASH

A DASH web application was developed in order for the HR team to be able to enter employee inputs to receive attrition predictions based on the inputs.

Implementation Details :

Loaded the trained Logistic regression model , scaler and label encoder from the artifacts directory.

The models’ predict\_proba() was used in to get the probability of attrition.

### 

### MODEL DEPLOYMENT

The web application and model was deployed on Render , where it will be hosted

A screenshot of a computer

AI-generated content may be incorrect.

PROJECT STRUCTURE

The files were saved according to the structure below

A screenshot of a computer program

AI-generated content may be incorrect.

## CHALLENGES FACED

Throughout the duration of the project there were a few areas that the team faced challenges in :

1. Handling the imbalanced data
2. Preprocess issues
3. Model optimisation
4. Model deployment

# Project links

### GitHub Repository

* Contains scripts , notebooks and artifacts
* Link : https://github.com/0101Keke/Employee\_Attrition\_Prediction

### Web Application

* Web app deployed on Render
* Link : [Employee Attrition Prediction](https://employee-attrition-prediction-9cwl.onrender.com/)

# CONCLUSION

The project set out to deploy a successful machine learning model to predict the employee attrition based on feature inputs , providing the HR team with the means of identifying at risk employees. Although a successful deployment was not achieved due to the challenges faced ,the project provided valuable insights into how intricate model development and deployment is. In retrospect the DASH application would have provided the HR team with an intuitive user-friendly interface that displayed predictive insights. This will give guidance for future improvements and give important considerations for projects moving forward.