Problem Statement

Objective

This assignment will provide hands-on experience building and evaluating a logistic regression model. You will explore data preprocessing techniques, feature selection and model implementation while gaining insights into the assumptions and working principles of logistic regression. By applying model evaluation and optimisation strategies, we will compare different approaches to enhance predictive performance. This will deepen your understanding of logistic regression and its role in making data-driven decisions across various real-world scenarios.

Business Objective

A mid-sized technology company wants to improve its understanding of employee retention to foster a loyal and committed workforce. While the organisation has traditionally focused on addressing turnover, it recognises the value of proactively identifying employees likely to stay and understanding the factors contributing to their loyalty.

In this assignment, you'll be building a logistic regression model to predict the likelihood of employee retention based on data such as demographic details, job satisfaction scores, performance metrics, and tenure. The aim is to provide the HR department with actionable insights to strengthen retention strategies, create a supportive work environment, and increase the overall stability and satisfaction of the workforce.

Dataset

Context

The data is stored as a CSV file and contains detailed information about various aspects of an employee's profile, including demographics, job-related features, and personal circumstances along with if the employee has left the company.

Content

The dataset consists of 74,610 rows and 24 columns, which are as follows:

- Employee ID: A unique identifier assigned to each employee.
- Age: The age of the employee, ranging from 18 to 60 years.
- Gender: The gender of the employee.
- Years at Company: The number of years the employee has been working at the company.
- Monthly Income: The monthly salary of the employee, in dollars.
- Job Role: The department or role the employee works in, encoded into categories such as Finance, Healthcare, Technology, Education, and Media.
- Work-Life Balance: The employee's perceived balance between work and personal life (Poor, Below Average, Good, Excellent).
- Job Satisfaction: The employee's satisfaction with their job (Very Low, Low, Medium, High).
- Performance Rating: The employee's performance rating (Low, Below Average, Average, High).
- Number of Promotions: The total number of promotions the employee has received.
- Overtime: Total number of overtime hours.
- Distance from Home: The distance between the employee's home and workplace, in miles.

- Education Level: The highest education level attained by the employee (High School, Associate Degree, Bachelor's Degree, Master's Degree, PhD).
- Marital Status: The marital status of the employee (Divorced, Married, Single).
- Number of Dependents: Number of dependents the employee has.
- Job Level: The job level of the employee (Entry, Mid, Senior).
- Company Size: The size of the company the employee works for (Small, Medium, Large).
- Company Tenure (In Months): The total number of years the employee has been working in the industry.
- Remote Work: Whether the employee works remotely (Yes or No).
- Leadership Opportunities: Whether the employee has leadership opportunities (Yes or No).
- Innovation Opportunities: Whether the employee has opportunities for innovation (Yes or No).
- Company Reputation: The employee's perception of the company's reputation (Very Poor, Poor, Good, Excellent).
- Employee Recognition: The level of recognition the employee receives(Very Low, Low, Medium, High).
- Attrition: Whether the employee has left the company.

Acknowledgements

This dataset is free and is publicly available at Kaggle.

Instructions

1. This is an ungraded assignment and requires no submission.
2. The programming language is Python.
3. You will be provided with the data set and a starter notebook. You have to perform all the tasks in the starter notebook.
4. The data will have inconsistencies and outliers; please handle them as you see fit and mention them in your report.
5. You are encouraged to search the web and consult AI tools for conceptual understanding.
6. You should prepare these two files:
   (a) an Interactive Python Notebook (.ipynb) that contains your code
   (b) a report Document (.pdf) that presents your visualisations, analysis, results, insights, and outcomes.
7. Mention all assumptions, if made, in the report.
8. The report should include the overall approach of the assignment, covering the problem statement, methodology, techniques used and key insights. Any graphs/plots you generate for analysis should also be attached to the report.

Results Expected from Learners

1 Report

Follow the instructions given below for creating the report document:

1. Provide a brief description of the problem statement and outline the step-by-step approach used in the assignment.
2. Provide insightful and well-labelled visualisations to support the analysis.
3. Summarise the final insights and their implications for employee retention strategies.
4. Propose data-driven recommendations to help the company improve employee retention.
5. Ensure the report is clear, concise, and uses business language suitable for stakeholders.

2 Starter Notebook

In the starter notebook, you will find headings, subheadings, and checkpoints stating the tasks you need to perform. The marks associated with each checkpoint will also be mentioned in the notebook. Keep in mind not to edit the cells with marking schemes and questions. You can find a brief description of the tasks below.

1. Data Understanding

Load the data and understand the basic statistical summary of the data.

2. Data Cleaning (15 marks)

2.1 Handle the missing values (10 marks)

2.2 Identify and handle redundant values within categorical columns (if any) (3 marks)

2.3 Drop redundant columns (2 marks)

3. Train - Validation Split (5 marks)

3.1 Split the data into train and validation with 70-30 ratio.

4. EDA on Training Data (20 marks)

4.1 Perform univariate analysis (6 marks)

4.2 Perform correlation analysis (4 marks)

4.3 Check class balance (2 marks)

4.4 Perform bivariate analysis (8 marks)

5. EDA on Validation Data (Optional)

5.1 Perform univariate analysis

5.2 Perform correlation analysis

5.3 Check class balance

5.4 Perform bivariate analysis

6. Feature Engineering (20 marks)

6.1 Dummy variable creation (15 marks)

6.2 Feature scaling (5 marks)

7. Model Building (40 marks)

7.1 Feature selection (5 Marks)

7.2 Building logistic regression model (20 marks)

7.3 Find the optimal cutoff (15 marks)

8. Prediction and Model Evaluation (30 marks)

8.1 Make predictions over validation set (15 marks)

8.2 Calculate accuracy of the model (2 marks)

8.3 Create confusion matrix and create variables for true positive, true negative, false positive and false negative (5 marks)

8.4 Calculate sensitivity and specificity (4 marks)

8.5 Calculate precision and recall (4 marks)

Rubrics

Consider the following criteria while preparing your solutions to the above tasks.

Table 1: Rubrics

Criteria	Meets expectations	Does not meet expectations
Data Cleaning	1. Missing values are handled correctly. 2. Redundant values within categorical columns are correctly identified and handled (if any). 3. Redundant columns are dropped correctly.	1. Missing values are handled inadequately. 2. Redundant values within categorical columns are incorrectly identified and handled (if any). 3. Redundant columns are not dropped correctly.
Train-Validation Split	1. Feature and target variables are defined correctly 2. Data is split into training and validation sets maintaining the ratio of 70:30.	1. Feature and target variables are not defined or are defined incorrectly 2. Data splitting is implemented incorrectly or not implemented at all.
EDA on training data	1. Performed univariate analysis by visualising the distribution of all numerical columns. 2. Performed correlation analysis by visualising a heatmap of the correlation matrix. 3. Visualised class distribution of the target variable. 4. Performed bivariate analysis by visualising the relationship between categorical columns and the target variable.	1. Failed to plot the distributions of numerical columns or created incomplete or inaccurate plots without meaningful insights. 2. Incorrectly visualised the heatmap or failed to interpret correlations. 3. Failed to visualise the class distribution of the target variable. 4. Failed to visualise the influence of categorical variables on the target variable. Provided incomplete or unclear visualisations and insights.

Continued on next page

Table 1: Rubrics (Continued)

Criteria	Meets expectations	Does not meet expectations
Feature Engineering	1. Created dummy variables for independent and dependent columns in both training and validation sets. 2. Applied feature scaling to numerical columns effectively by scaling the features.	1. Failed to identify or create appropriate dummy variables for independent and dependent columns in both training and validation sets. 2. Failed to apply or incorrectly performed feature scaling, leading to inconsistent data ranges.
Model Building	1. Selected the most important features using Recursive Feature Elimination (RFE). 2. Built a logistic regression model using the selected features, evaluated multicollinearity with p-values and VIFs, made predictions, and assessed model performance. 3. Found the optimal cutoff by plotting the ROC curve, visualising trade-offs between sensitivity and specificity, precision and recall, and evaluated the final prediction with optimal cutoff.	1. Failed to select appropriate features using Recursive Feature Elimination(RFE). 2. Failed to correctly build the model using selected features and evaluated or interpreted the performance metrics incorrectly. 3. Failed to identify the optimal cutoff or omitted key evaluations like plotting curves and calculated necessary performance metrics.
Prediction and Model Evaluation	1. Prediction were made on validation data by selecting the relevant features. 2. Evaluated the model's performance correctly using given evaluation metrics.	1. Failed to select relevant features to make predictions on validation data. 2. Failed to evaluate the model using correct evaluation metrics or inaccurately.

Continued on next page

Table 1: Rubrics (Continued)

Criteria	Meets expectations	Does not meet expectations
Report and Recommendations	1. The report has a clear structure, is not too long, and explains the most important results concisely in simple language. 2. The recommendations to solve the problem are realistic, actionable and coherent with the analysis. 3. The report includes visualisations and insights derived from them. 4. If any assumptions are made, they are stated clearly.	1. The report lacks structure, is too long or does not put emphasis on the important observations. The language used is complicated for business people to understand. 2. The recommendations to solve the problem are either unrealistic, non-actionable or incoherent with the analysis. 3. The report is missing visualisations or fails to provide meaningful insights. 4. Assumptions made, if any, are not stated clearly.
Conciseness and readability of the code	1. The code is concise and syntactically correct. Wherever appropriate, built-in functions and standard libraries are used instead of writing long code (if-else statements, loops, etc.). 2. Custom functions are used to perform repetitive tasks. 3. The code is readable with appropriately named variables and detailed comments are written wherever necessary.	1. Long and complex code is used instead of shorter built-in functions. 2. Custom functions are not used to perform repetitive tasks resulting in the same piece of code being repeated multiple times. 3. Code readability is poor because of vaguely named variables or lack of comments wherever necessary.