Report predicting-employee-retention-starter

thang.truong

This project was completed solely by me due to difficulties in contacting other team members.

Overview

- Employee attrition is a critical challenge for organizations aiming to retain talent and maintain workforce stability. This report presents a logistic regression model designed to predict employee attrition using key features such as **age, monthly income, years at the company, and job role**.
- The goal is to analyze patterns and identify employees at risk of leaving, enabling proactive retention strategies.

Data Preprocessing

2.1. Dataset Description

- The dataset, sourced from GitHub, includes various employee attributes such as personal demographics, work history, and attrition status.

2.2. Feature Selection

- The model uses the following features:
  - **Age**
  - **Monthly Income**
  - **Years at Company**
  - **Job Role**

2.3. Handling Categorical Data

- Categorical variables such as *Job Role* were converted into numeric values using **one-hot encoding** to ensure compatibility with logistic regression.

2.4. Splitting Data

- The dataset was divided into **training (80%)** and **validation (20%)** sets for model evaluation.

Model Training

A logistic regression model was trained using **scikit-learn** with 1,000 iterations to ensure convergence.

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression(max_iter=1000)

model.fit(X_train, y_train)
```

Predictions on the Validation Set

y_validation_pred = model.predict_proba(X_val)[:, 1] # Get probabilities

To classify predictions based on an **optimal cutoff threshold (e.g., 0.5)**:

optimal_cutoff = 0.5

results_df['final_prediction'] = (results_df['Predicted_Probability'] >= optimal_cutoff).astype(int)