Employee Retention Analysis

Developing a Logistic Regression Model for Predictive Insights

Introduction

This presentation outlines the development of a Logistic Regression model aimed at analyzing employee retention in a mid-sized technology company. The model will leverage various employee data to predict binary outcomes regarding retention, thereby aiding HR strategies.

01

Logistic Regression Overview

Assumptions of Logistic Regression

Logistic regression operates under several assumptions, including:

1. The dependent variable is binary.

2. Observations are independent of each other.

3. There is a linear relationship between the independent variables and the log odds of the dependent variable.
4. No multicollinearity exists among the independent variables.
Understanding these assumptions is critical in ensuring the validity of the model's predictions.

Binary Outcome Predictions

Binary outcome predictions in logistic regression involve estimating the probability of a specific outcome occurring based on input features. By applying the logistic function, the model transforms linear combinations of input variables into a probability score ranging from 0 to 1. If the predicted probability exceeds a defined threshold (commonly 0.5), the model classifies the outcome as one category (e.g., retention); otherwise, it classifies it as the other category (e.g., attrition). This approach allows organizations to identify employees at risk of leaving and implement appropriate retention strategies.

02

Model Development Process

Data Preparation and Cleaning

Effective data preparation is critical to building a reliable logistic regression model. This process involves:

Data Understanding: Familiarizing with the dataset, including the 24 columns and 74610 rows, and the information each column presents.

Handling Missing Values: Distance from Home and Company Tenure (In Months) columns has a 3 % null values so we remove these values from dataset.

Minmaxsclaer: Adjusting numerical features to a common scale, particularly when features vary significantly in range.

Feature Engineering Techniques

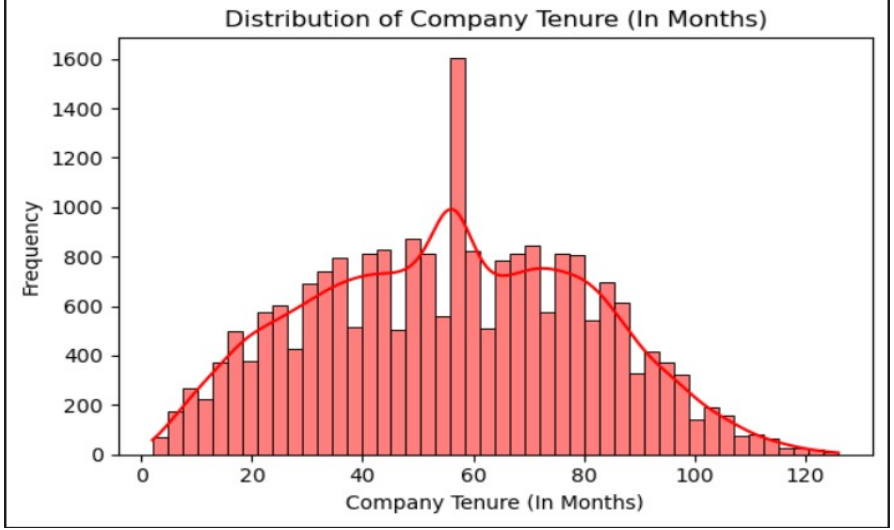
Feature engineering plays a key role in enhancing the predictive power of the logistic regression model. Some techniques include:

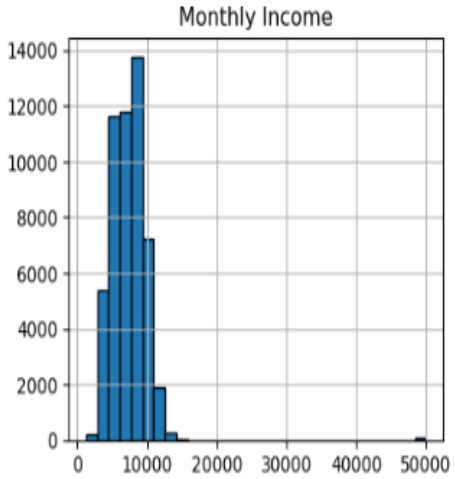
Train and test:Using a train_test_split we split over data in two parts training and testing training 70% data and testing 30% data.

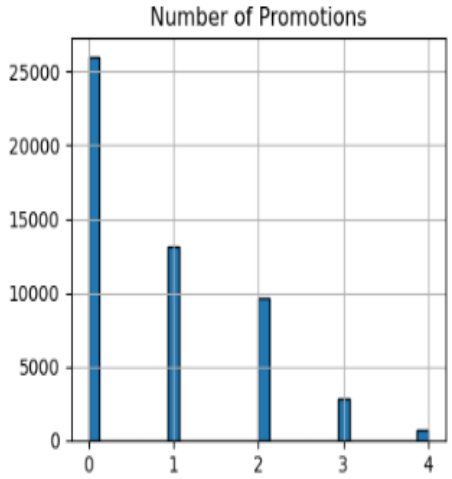
Encoding Categorical Variables: Converting categorical variables, such as Job Role and Education Level, into numerical format using a get_dummies method.

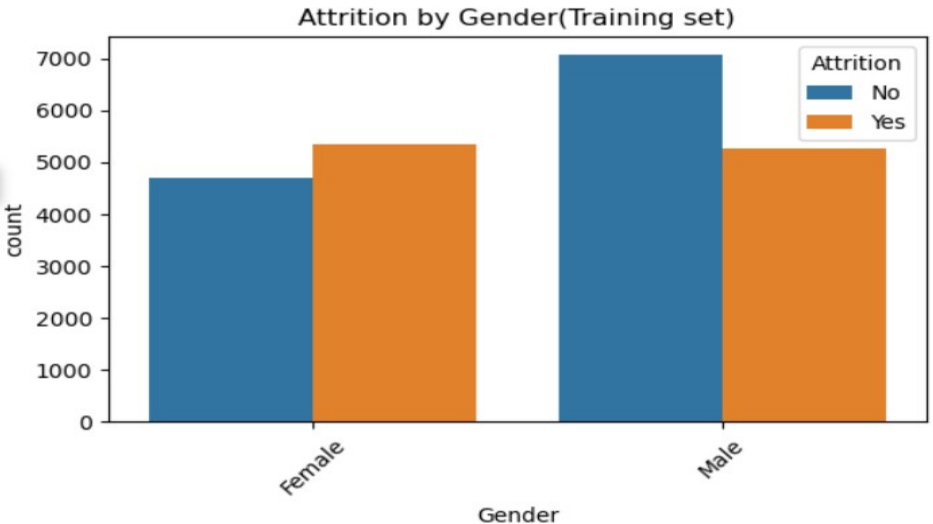
Scaling:Scaling the data x_train and x_test using minmaxscaler in fix range.

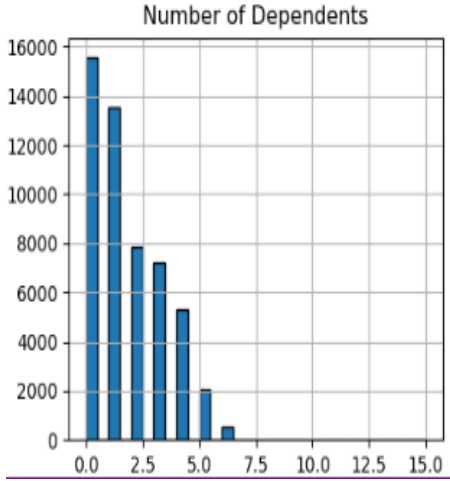
EDA: Using a eda we extract some meaning full ensights from data.

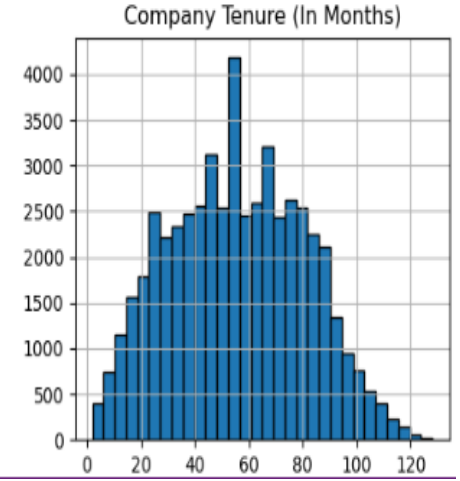

Distribution of Company Tenure (In Months)
Frequency
Company Tenure (In Months)


Monthly Income

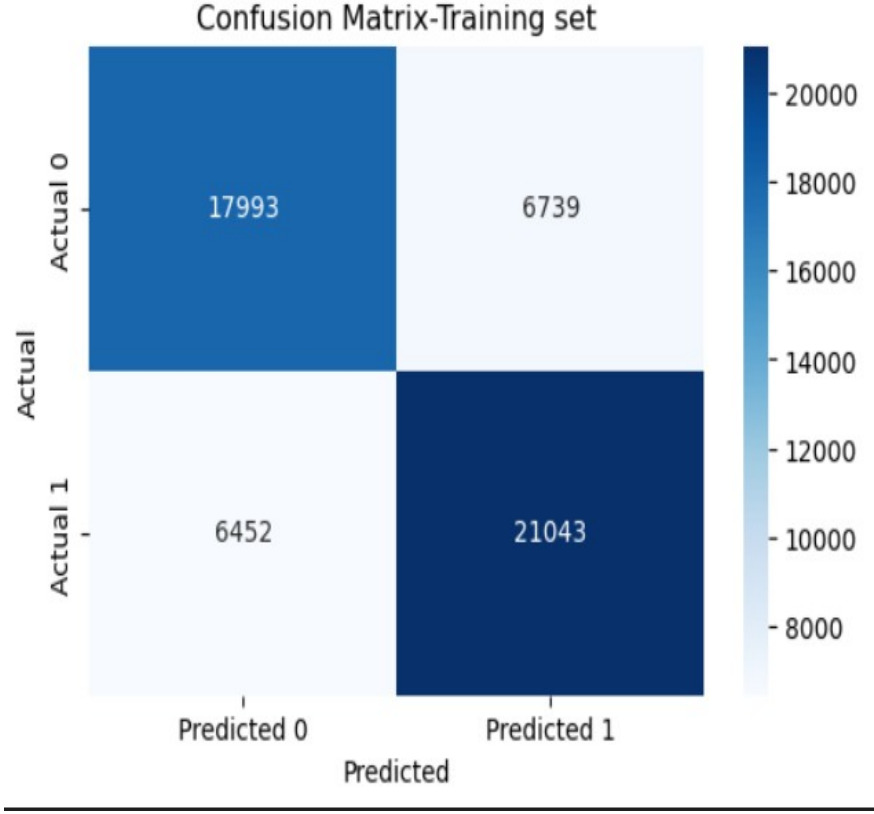

Number of Promotions


Attrition by Gender(Training set)
Attrition
No
Yes
count
Female
Male
Gender


Number of Dependents


Company Tenure (In Months)


Target class Distribution in training set
12000
10000
8000
6000
4000
2000
0
count
No
Yes
Target class


Confusion Matrix-Training set
Actual 0
Actual 1
Actual
17993
6739
6452
21043
Predicted 0
Predicted 1
Predicted
20000
18000
16000
14000
12000
10000
8000

Model Building and Evaluation Metrics

We use a Logistic Regression algo for model building.

Recursive Feature Elimination :Since we had many input features, we applied a feature selection technique called RFE to select top15 features.

Accuracy: The proportion of true results (both true positives and true negatives) among the total number of cases examined.The overall Training accuracy 0.7474

Precision and Recall: Precision measures the accuracy of positive predictions, while recall (sensitivity) measures the ability to identify all relevant instances.In this model the overall

Precision 0.7574

And recall 0.7653

Conclusions

By carefully preparing the data and using Logistic Regression with feature selection through Recursive Feature Elimination (RFE), we built a reliable classification model. We started with a dataset containing 24 features and 74,610 records, handled missing values, standardized the data, encoded categorical variables, and selected the top 15 most important features. The model achieved **74% accuracy** , indicating consistent performance. Additionally, the **precision is** 75.74% **and recall values is** 76.53 **%** showed that the model is fairly balanced in identifying both positive and negative cases. These results demonstrate that with proper preprocessing and feature selection, Logistic Regression can provide strong predictive performance in real-world classification problems.