Predicting Employee Retention

Logistic Regression Assignment

By Anshaj Singh

Contents :

1. Problem Statement ………………………………………………………………………….. 3
2. Approach …………………………………………………………………………………….. 3
3. Techniques & Libraries Used ..……………………………………………………………… 3
4. Key Insights - Technical ……………………………………………………………………….. 4
5. Key Insights - Business Specific ……………………………………………………………….. 4
6. Plots/Graphs ……………………………………………………………………………………… 5

Problem Statement :

Predict whether an employee will stay or leave the company using demographic, job, compensation, and satisfaction/work-life features. Use logistic regression to build a predictive model; interpret drivers of attrition and provide insights that HR can act on.

Approach :

1. Data understanding
    - Loaded the CSV and inspected the data to understand sizes and data types
    - Checked categorical unique values and the target (Attrition) distribution
2. Data Cleaning
    - Checked for missing values
    - Handled redundant categorical values and normalized columns where needed.
    - Dropped redundant columns
3. Train-Validation Split
    - Split dataset into train (70%) and validation (30%) sets using train_test_split with stratification on the target to preserve class balance
4. EDA on Training Data
    - Performed Univariate Analysis
    - Performed Co-relation Analysis
    - Checked Class Balance
    - Performed Bivariate Analysis
5. Feature Engineering
    - Created dummy variables for categorical columns using pd.get_dummies
    - Scaled numeric features
6. Feature Selection & Model Building
    - Used RFE with a logistic regression estimator to select a subset of important features
    - Building Logistic Regression Model
    - Evaluated multicollinearity using VIF and examined p-values/ coefficients to assess significance
7. Cutoff Selection
    - Computed predicted probabilities, plotted metrics (accuracy, specificity & sensitivity) across thresholds
    - Selected an optimal cutoff by balancing Sensitivity and Specificity (optimal cutoff- 0.52)
8. Prediction & Evaluation
    - Applied the final model to the validation set.
    - Computed predictions (using chosen cutoff) and prediction probabilities
    - Calculated accuracy, confusion matrix, TP/TN/FP/FN, sensitivity, specificity, precision, and recall.
    - Created visualizations for confusion matrix, ROC, cutoff vs metrics, etc
9. Conclusion
    - Generated summary outputs and recommendations for HR action based on model insights.

Techniques & Libraries Used :

1. Data manipulation: pandas, numpy
2. Plotting / EDA: matplotlib, seaborn
3. Train/test split: sklearn.model_selection.train_test_split
4. Encoding: pd.get_dummies
5. Scaling: StandardScaler/ MinMaxScaler
6. Feature selection: sklearn.feature_selection.RFE
7. Modeling: sklearn.linear_model.LogisticRegression
8. Model diagnostics: p-values (, VIF (

9. Model evaluation: roc_curve, auc, confusion_matrix, precision_score, recall_score, accuracy_score

Key Insights - Technical :

1. Dataset shape: 74,610 rows × 24 columns
2. Only two columns have missing values - ‘Distance From Home’ (2.56%) & ‘Company Tenure’ (3.23%). Since these columns have less than 5% missing values, therefore the rows with missing values were dropped. % of the remaining dataset = 94.67%
3. Target distribution: Stayed = 39,191 (52.53%); Left = 35,419 (47.47%)
4. Optimal cutoff = 0.52
5. Performance at the chosen cutoff (validation):
    - Accuracy ≈ 0.7372
    - Sensitivity ≈ 0.7394
    - Specificity ≈ 0.7348
6. RFE was used to select the top features and VIF/p-values were checked (multicollinearity examined)
7. Markdown explanations / summary are written as needed in the notebook.

Key Insights - Business Specific :

From the logistic regression coefficients, several patterns emerged as strong predictors of attrition:

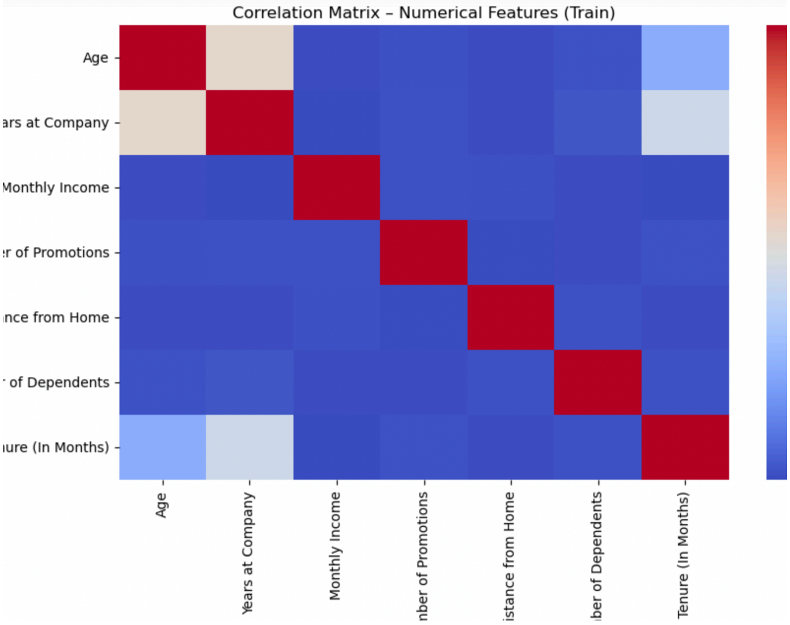
1. Poor or fair work–life balance significantly increases attrition risk
2. Low job satisfaction is a key driver of leaving
3. Employees who work overtime have higher attrition probability
4. Lower performance ratings correlate with leaving
5. Single employees show higher mobility compared to married ones
6. Higher job levels (Mid, Senior) and PhD education levels are associated with higher retention
7. Remote work availability strongly increases retention

These findings provide important direction for HR and leadership teams. Improving work–life balance, monitoring workload, focusing on early-career employees, and expanding flexible work policies could positively influence retention.

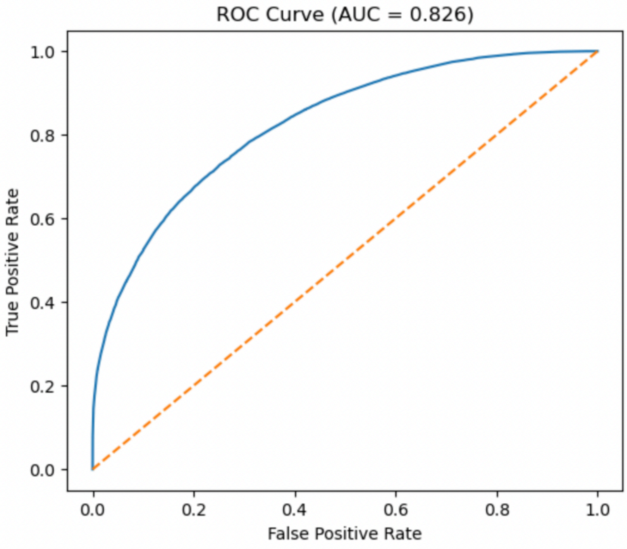
Overall, the model is interpretable, statistically valid, and gives actionable insights that can support data-driven employee retention strategies.

Plots/Graphs :

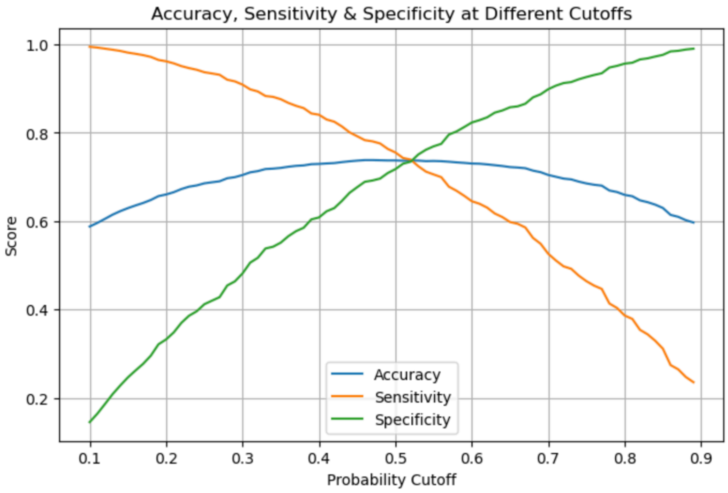
1. Co-relation Matrix on Numerical Features on Training Dataset



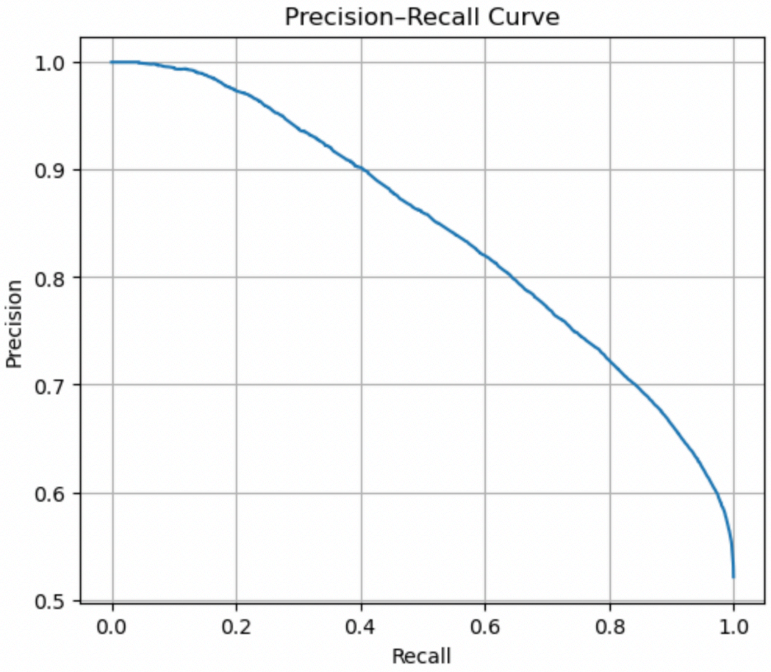
2. ROC Curve



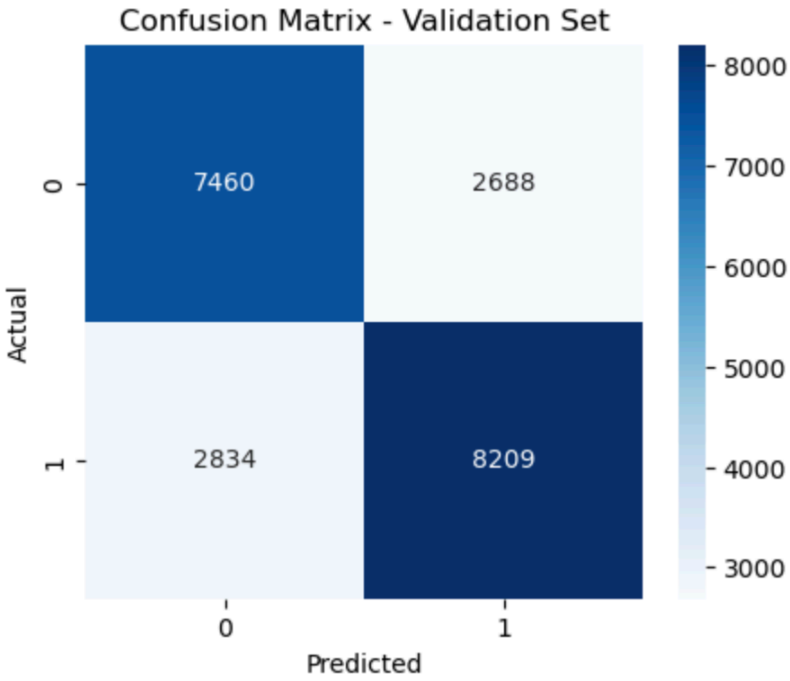
3. Accuracy, Sensitivity & Specificity At Different Cut-offs



4. Precision-Recall Curve



5. Confusion Matrix - Validation Set



Caption