Understanding The Concept of Bias : 01

Problem Statement (As Defined By The Client)

Client Requirement

- A Large HR Consulting Firm Wants To Develop A Predictive Model That Accurately Forecasts Employee Attrition Based on Various Factors Such As Job Satisfaction, Salary, Department, Work-Life Balance, And Years At The Company
- The Existing Predictive Model Has Been Observed To Underestimate The Actual Attrition Rates For Departments With High Job Dissatisfaction And Work-Life Imbalance
- The Client Has Provided Historical Employee Data, Including Actual Attrition Status And Various Attributes Such As Employee Age, Salary, Job Role, Job Satisfaction, Years At The Company, And Performance Rating
- The Client Wishes To Understand Why The Current Model Deviates From Actual Observations And How A New Model Can Be Developed To Provide A More Accurate Forecast of Employee Attrition
- The Client Seeks To Identify The Shortcomings of The Existing Model And Determine An Improved Model That Minimizes Prediction Errors, Thereby Helping in Better Workforce Planning And Retention Strategies

Specific Goals

1. Identify And Explain The Trend Differences Between Actual And Predicted Employee Attrition Using Visual Representations
2. Evaluate The Accuracy of The Existing Logistic Regression Model For Predicting Employee Attrition And Identify The Reasons For High Variance And Underfitting
3. Propose A Revised Model That Fits The Employee Data More Accurately

Problem Statement (As Defined By The Data Scientist)

Title

Evaluating And Improving The Predictive Accuracy of Employee Attrition Using Machine Learning Techniques

Objective

- To Analyze And Improve The Existing Predictive Model For Employee Attrition That Currently Exhibits High Variance And Underfitting
- The Project Aims To Identify Discrepancies Between Actual And Predicted Employee Attrition And Propose An Enhanced Model That Minimizes Prediction Error, Thus Improving The Alignment of Predicted Attrition With Actual Employee Exits

Problem Description

- The Current Predictive Model Used By The HR Consulting Firm is Based on A Logistic Regression Formula

  Predicted Attrition Probability = $1 / (1 + e^{(-\text{LinearCombination})})$

  LinearCombination = BaseCoefficient + β1*JobSatisfaction + β2*YearsAtCompany + β3*PerformanceRating

- However, When Compared To The Actual Attrition, It Appears That The Predicted Values Show Considerable Underestimation For Departments With High Job Dissatisfaction And Work-Life Imbalance

  Actual Attrition = AttritionStatus [0 or 1]

Understanding The Concept of Bias : 01

- The Linear Nature of The Current Model Does Not Accurately Capture The Interaction Between Job Satisfaction, Work-Life Balance, And Years At The Company. This Results in A Model With High Variance And Underfitting, Missing The Mark on High-Attrition Departments, Which is Detrimental For Workforce Planning

Remarks on The Current Model

- The Predicted Values Consistently Underestimate The Attrition Rate For Departments With Lower Job Satisfaction And Poorer Work-Life Balance
- The Discrepancy Suggests That The Existing Logistic Regression Model is Insufficient To Capture The True Relationship Between Employee Attributes And Attrition
- The High Variance Leads To Significant Underfitting, Indicating The Need For A More Complex Model That Can Capture Interactions Between Different Employee Attributes

Approach

1. **Data Analysis**

   Perform Exploratory Data Analysis (EDA) To Understand The Distribution of Employee Demographics, Job Satisfaction, Work-Life Balance, And Other Attributes Along With The Corresponding Actual Attrition Status

2. **Model Evaluation**

   Plot And Analyze The Actual Versus Predicted Attrition To Visualize And Quantify The Variance

3. **Model Redesign**

   Implement And Test Alternative Models (e.g., Decision Trees OR Random Forests) To Capture The Non-Linear Relationship Between Employee Attributes And Attrition

4. **Model Validation**

   Evaluate The Revised Model's Performance Using Appropriate Metrics Such As

   - Mean Absolute Error (MAE)
   - Root Mean Squared Error (RMSE)
   - R-Squared Value.

Expected Outcome

- An Improved Predictive Model That Significantly Reduces Variance And Error, Accurately Predicting Employee Attrition Across Different Departments And Work-Life Balances
- This Will Provide The Client With A More Reliable Tool For Estimating Employee Exits, Leading To Better Workforce Planning And Retention Strategies

Sample Data

Employee ID	Age	Gender	Department	Job Satisfaction	Salary	Work Life Balance	Years At Company	Performance Rating	Attrition (Actual)
101	34	Male	Sales	2	65000	3	5	4	0
102	29	Female	HR	4	58000	4	3	3	0
103	42	Male	IT	3	85000	2	8	5	1
104	25	Female	Marketing	1	47000	1	2	3	1
105	37	Male	Finance	2	74000	3	6	4	0
106	31	Female	Operations	3	68000	4	4	4	0

107	45	Male	IT	2	90000	2	10	5	1
108	28	Female	Sales	4	62000	3	3	3	0
109	39	Male	HR	1	56000	1	5	4	1
110	33	Female	Marketing	3	50000	3	4	4	0
111	41	Male	Operations	2	72000	2	7	5	0
112	26	Female	Sales	1	45000	1	2	3	1
113	38	Male	Finance	3	78000	3	6	4	0
114	30	Female	HR	4	60000	4	4	4	0
115	44	Male	IT	1	89000	1	9	5	1

Variables Assumed

1. **Base Salary**: $50,000 : The Average Starting Salary For Employees in The Organization
2. **Base Job Satisfaction**: 3.0 : A Baseline Level of Job Satisfaction on A Scale of 1 To 5
3. **Base Work-Life Balance**: 3.0 : A Baseline Level of Work-Life Balance on A Scale of 1 To 5
4. **Salary Coefficient**: 0.005 : The Increase in Attrition Risk Per Additional Dollar in Salary
5. **Job Satisfaction Coefficient**: -0.7 : The Decrease in Attrition Risk For Each Point Increase in Job Satisfaction
6. **Work-Life Balance Coefficient**: -0.6 : The Decrease in Attrition Risk For Each Point Increase in Work-Life Balance
7. **Years At Company Coefficient**: -0.1 : The Decrease in Attrition Risk For Each Additional Year At The Company
8. **Performance Rating Coefficient**: -0.5 : The Decrease in Attrition Risk For Each Point Increase in Performance Rating
9. **Age Factor**: 0.02 : The Increase in Attrition Risk For Each Additional Year of Age
10. **High Job Dissatisfaction Threshold**: 2.5 : Job Satisfaction Score Below This Threshold is Considered High Risk For Attrition
11. **Low Work-Life Balance Threshold**: 2.5 : Work-Life Balance Score Below This Threshold is Considered High Risk For Attrition

Assumptions

- The Coefficients Are Determined Based on Historical Data Analysis And Regression Modeling
- The Variables Interact Linearly, Meaning Each Factor Contributes Independently To The Attrition Probability