Advanced Data Analysis Report

Overview

This report summarizes the advanced data analysis performed on the employee attrition dataset. It includes statistical tests, feature selection insights, and recommendations for further steps in the project.

---

1. Statistical Tests

1.1 t-Tests

Objective: Compare the means of numerical variables between employees who left (Attrition=Yes) and those who stayed (Attrition=No).

Significant Variables (p-value < 0.05):

1. **Age**: Employees who left tend to be younger than those who stayed (τ-statistic: -5.83, p ≈ 1.38 × 10^-8).
2. **DailyRate**: A minor but significant difference was observed (p ≈ 0.03).
3. **MonthlyIncome**: Lower monthly income is significantly associated with attrition (τ-statistic: -7.48, p ≈ 4.43 × 10^-13).
4. **JobLevel**: Higher job levels correlate with lower attrition (τ-statistic: -7.39, p ≈ 9.84 × 10^-13).
5. **YearsAtCompany**: Employees with shorter tenures are more likely to leave (τ-statistic: -5.28, p ≈ 2.29 × 10^-7).

Interpretation: These variables should be prioritized for predictive modeling as they exhibit strong differences between attrition groups.

1.2 Chi-Squared Tests

Objective: Assess the relationship between categorical variables and attrition.

Significant Variables (p-value < 0.05):

1. **BusinessTravel**: Employees with frequent travel are more likely to leave (χ^2-statistic: 24.18, p ≈ 5.61 × 10^-6).
2. **OverTime**: Employees working overtime show higher attrition (χ^2-statistic: 87.56, p ≈ 8.16 × 10^-21).
3. **MaritalStatus**: Single employees have higher attrition rates (χ^2-statistic: 46.16, p ≈ 9.46 × 10^-11).

4. **JobRole**: Significant variations in attrition exist across roles (χ^2-statistic: 86.19, p ≈ 2.75 × 10^-15).

Interpretation: These categorical variables are highly associated with attrition and should be encoded as features for modeling.

1.3 ANOVA

Objective: Analyze numerical variables across multiple categorical groups (e.g., JobRole).

Key Findings:

- **MonthlyIncome**: Significant differences in income levels across job roles (F-statistic: High, $p < 0.05$).
- **YearsAtCompany**: Varies significantly by job role, with implications for tenure categories.

Interpretation: These insights can guide the creation of interaction features (e.g., Income-to-Role Ratio).

---

2. Feature Engineering

2.1 Newly Created Features:

1. **Tenure Categories**:
    - Short-Term: < 3 years
    - Medium-Term: 3-7 years
    - Long-Term: > 7 years
2. **Salary Bands**:
    - Low: <$3000
    - Medium: $3000-$7000
    - High: >$7000
3. **Interaction Features**:
    - Performance-to-Salary Ratio
    - YearsSinceLastPromotion-to-YearsAtCompany

2.2 Transformations:

1. **Normalization**: Applied to numerical variables (e.g., MonthlyIncome, Age).
2. **Encoding**: One-hot encoding for categorical variables (e.g., JobRole, MaritalStatus).

3. Visualization Insights

3.1 Attrition Trends:

- **Heatmaps**: Highlighted strong correlations between OverTime and attrition.
- **Bar Charts**: Showed higher attrition rates for singles and employees in lower-income bands.

3.2 Recommendations:

- Use interactive dashboards to present trends (e.g., Tableau, Plotly).

4. Recommendations for Modeling

1. Prioritize significant variables identified in statistical tests.
2. Include engineered features like tenure categories and interaction terms.
3. Address class imbalance using techniques like SMOTE or undersampling.