Enhancing Human Resources Insights through Advanced Analytics

Nguyen Thanh Hoa
Student ID: SE183091

July 16, 2024

1 Objective

This project aims to leverage data analytics and machine learning to gain a deeper understanding of Human Resources (HR) operations. This understanding can help optimize workforce management, improve employee satisfaction, and enhance overall organizational effectiveness. The project focuses on analyzing HR metrics such as demographics, performance scores, and compensation.

2 Specific Objectives

- Improve workforce management.

- Increase employee satisfaction.

- Optimize compensation strategies.

- Provide actionable predictive insights.

3 Methodology

The project is carried out in the following steps:

1. **Data Preparation and Cleaning:**

   - Handling Missing Values:
     - Replace missing values for numerical attributes with the mean.
     - Replace missing values for categorical attributes with the most frequent value.
   - Feature Engineering:
     - Create age groups.
   - Data Encoding and Normalization:

- Encode categorical attributes using OneHotEncoder.
- Normalize numerical attributes using StandardScaler.
- Outlier Handling using IQR.

2. **Exploratory Data Analysis (EDA):**

- Calculate descriptive statistics.
- Visualize data distributions:
  - Distribution of attrition status, attrition rate by department...
  - Correlation heatmap.
- Identify patterns and support decision-making.

3. **Data Visualization:**

- Distribution Plots:
  - Attrition status.
  - Employees by job role.
  - Attrition rate by department.
  - Monthly income by job role.
- Correlation heatmap.

4. **Model Building and Evaluation:**

- Objective: Predict employee attrition and performance.
- Models Used:
  - **Classification:**
    * Logistic Regression: Classify attrition status.
    * Random Forest: Classify attrition status.
    * XGBoost: Classify attrition status.
    * LightGBM: Classify attrition status.
  - **Regression:**
    * Regression: Predict PerformanceRating.
    * Gradient Boosting: Predict PerformanceRating.

5. **Evaluation, Conclusion, and Recommendations:**

- Evaluate model results.
- Draw conclusions.
- Propose future directions.

4 Results

4.1 Exploratory Data Analysis

- Attrition Rate: The chart shows that the attrition rate ("Yes") accounts for a significant portion, which may be a sign to investigate the reasons and develop employee retention solutions.

- Employee Distribution: The number of employees is relatively evenly distributed across most positions and departments within the company.

- Compensation Policy: Relatively fair, however, there is still income differentiation among individuals in the same job position. This suggests that the company may be applying a compensation policy based on individual competence and experience.

- Long-Term Employees: Typically have higher levels, income, and age.

4.2 Model Prediction Results

- **Classification:** Logistic Regression, RandomForestClassifier, XGBoost, and LightGBM models have relatively ineffective prediction rates with an Accuracy ratio fluctuating around 50%.

- **Regression:** The Gradient Boosting and Random Forest Regressor models have MSE, RMSE, and MAE approximations ranging from 0.99 to 1.24, indicating a large prediction error. Notably, the R2 is close to 0 and negative, suggesting that the model predicts less effectively than simply using the average value. Gradient Boosting performs more stably than Random Forest Regressor.

4.3 Feature Importance

- MonthlyIncome is the most important factor in predicting attrition.

- Other factors have relatively equal importance including: DailyRate, DistanceFromHome, PercentSalaryHike, TotalWorkingYears, Age, YearsAtCompany, YearsWithCurrManager, YearsInCurrentRole, NumCompaniesWorked.

4.4 SHAP Value Analysis

- The interaction between age and daily rate has a weak impact on prediction results.

- Age has a significant impact on prediction results, while daily rate has an unclear impact.

5 Conclusion and Recommendations

The project successfully analyzed HR data and built predictive models. However, the effectiveness of the predictive models is not high. Here are some recommendations to improve the project:

- **Collect More Data:** More comprehensive data can help improve model effectiveness.

- **Experiment with Different Models:** Other machine learning models can be used to find a more suitable model.

- **Tune Hyperparameters:** Hyperparameters of the model need to be tuned to achieve better performance.

- **Deeper Analysis of Features:** A deeper analysis of the importance of features can help select better features for the model.

6 Source Code

The project source code is stored at `https://github.com/Jikay-070203/Project_Business-Case_and_HR-Analytics`.