Enhancing Human Resources Insights through Advanced Analytics (Nâng cao hiểu biết sâu sắc về nguồn nhân lực thông qua phân tích nâng cao)

Nguyễn Thanh Hòa
Mã số sinh viên: SE183091

16 tháng 7, 2024

1 Mục tiêu

Dự án này hướng đến việc sử dụng phân tích dữ liệu và học máy để hiểu sâu hơn về hoạt động nhân sự (HR). Từ đó, dự án có thể giúp tối ưu hóa việc quản lý lực lượng lao động, cải thiện sự hài lòng của nhân viên và nâng cao hiệu quả tổng thể của tổ chức. Trọng tâm của dự án là phân tích các số liệu nhân sự như nhân khẩu học, điểm hiệu suất và lương thưởng.

2 Mục tiêu cụ thể

- Cải thiện quản lý lực lượng lao động.
- Tăng sự hài lòng của nhân viên.
- Tối ưu hóa chiến lược lương thưởng.
- Cung cấp thông tin chi tiết có thể dự đoán.

3 Phương pháp

Dự án được thực hiện theo các bước sau:

1. **Chuẩn bị dữ liệu và làm sạch dữ liệu:**
   - Xử lý giá trị thiếu:
     - Thay thế giá trị thiếu cho các thuộc tính số bằng giá trị trung bình.

  - Thay thế giá trị thiếu cho các thuộc tính hạng mục bằng giá trị xuất hiện nhiều nhất.
- Khai thác đặc trưng:
  - Tạo nhóm tuổi.
- Mã hóa, chuẩn hóa dữ liệu:
  - Mã hóa các thuộc tính hạng mục bằng OneHotEncoder.
  - Chuẩn hóa các thuộc tính số bằng StandardScaler.
- Xử lý ngoại lệ bằng cách sử dụng IQR.

2. **Phân tích dữ liệu thăm dò (EDA):**
   - Tính toán thống kê mô tả.
   - Trực quan hóa phân bố dữ liệu:
     - Phân bố tình trạng nghỉ việc, tỷ lệ nghỉ việc theo phòng ban...
     - Biểu đồ nhiệt tương quan.
   - Từ đó, phát hiện và hỗ trợ đưa ra quyết định.
3. **Trực quan hóa dữ liệu:**
   - Biểu đồ phân bố:
     - Tình trạng nghỉ việc.
     - Nhân viên theo chức vụ.
     - Tỷ lệ nghỉ việc theo phòng ban.
     - Thu nhập hàng tháng theo vai trò công việc.
   - Biểu đồ nhiệt tương quan.
4. **Xây dựng và đánh giá mô hình dự đoán:**
   - Mục tiêu: Dự đoán mức độ tiêu hao và hiệu suất của nhân viên.
   - Các mô hình được sử dụng:
     - **Phân loại:**
       * Logistic Regression: Phân loại tình trạng nghỉ việc (Attrition).
       * Random Forest: Phân loại tình trạng nghỉ việc.
       * XGBoost: Phân loại tình trạng nghỉ việc.
       * LightGBM: Phân loại tình trạng nghỉ việc.
     - **Hồi quy:**
       * Regression: Dự đoán PerformanceRating (Đánh giá hiệu suất).
       * Gradient Boosting: Dự đoán PerformanceRating.
5. **Đánh giá kết luận và đưa ra đề xuất:**
   - Đánh giá kết quả của mô hình.
   - Đưa ra kết luận.
   - Đề xuất hướng phát triển.

4 Kết quả

4.1 Phân tích dữ liệu thăm dò

- Tỷ lệ nghỉ việc: Biểu đồ cho thấy tỷ lệ nghỉ việc ("Yes") chiếm một phần đáng kể, có thể là dấu hiệu cần quan tâm để tìm hiểu nguyên nhân và có giải pháp giữ chân nhân viên.
- Phân bố nhân viên: Số lượng nhân viên phân bố tương đối đồng đều ở hầu hết các vị trí và phòng ban trong công ty.
- Chính sách lương thưởng: Tương đối công bằng, tuy nhiên vẫn có sự phân hóa thu nhập giữa các cá nhân cùng vị trí công việc. Điều này cho thấy công ty có thể đang áp dụng chính sách lương thưởng dựa trên năng lực và kinh nghiệm cá nhân.
- Nhân viên làm việc lâu năm: Thường có cấp bậc, thu nhập và tuổi đời cao hơn.

4.2 Kết quả mô hình dự đoán

- **Phân loại:** Các mô hình Logistic Regression, RandomForestClassifier, XGBoost, LightGBM có tỉ lệ dự đoán tương đối chưa hiệu quả với tỉ lệ Accuracy khoảng dao động ở mức 50%.
- **Hồi quy:** Hai mô hình Gradient Boosting và Random Forest Regressor có MSE, RMSE, MAE xấp xỉ từ 0.99 tới 1.24 cho thấy sai số dự đoán lớn. Đặc biệt đáng chú ý là R2 gần bằng 0 và âm cho thấy mô hình dự đoán kém hiệu quả hơn so với việc chỉ sử dụng giá trị trung bình. Gradient Boosting dự đoán ổn định hơn Random Forest Regressor.

4.3 Tầm quan trọng của đặc trưng

- MonthlyIncome (thu nhập hàng tháng) là yếu tố quan trọng nhất trong việc dự đoán nghỉ việc.
- Các yếu tố khác có tầm quan trọng tương đối đồng đều bao gồm: DailyRate, DistanceFromHome, PercentSalaryHike, TotalWorkingYears, Age, YearsAtCompany, YearsWithCurrManager, YearsInCurrentRole, NumCompaniesWorked.

4.4 Phân tích giá trị SHAP

- Tương tác giữa tuổi và lương ngày có ảnh hưởng yếu đến kết quả dự đoán.
- Tuổi có ảnh hưởng đáng kể đến kết quả dự đoán, trong khi lương ngày có ảnh hưởng không rõ ràng.

5 Kết luận và đề xuất

Dự án đã thành công trong việc phân tích dữ liệu nhân sự và xây dựng mô hình dự đoán. Tuy nhiên, hiệu quả của mô hình dự đoán chưa cao. Dưới đây là một số đề xuất để cải thiện dự án:

- **Thu thập thêm dữ liệu:** Dữ liệu phong phú hơn có thể giúp cải thiện hiệu quả của mô hình.
- **Thử nghiệm các mô hình khác:** Có thể sử dụng các mô hình học máy khác để tìm kiếm mô hình phù hợp hơn.
- **Tinh chỉnh siêu tham số:** Cần tinh chỉnh siêu tham số của mô hình để đạt được hiệu quả tốt hơn.
- **Phân tích sâu hơn về các đặc trưng:** Phân tích sâu hơn về tầm quan trọng của các đặc trưng có thể giúp lựa chọn đặc trưng tốt hơn cho mô hình.

6 Mã nguồn

Mã nguồn của dự án được lưu trữ tại `https://github.com/Jikay-070203/Project_Business-Case_and_HR-Analytics`.