Machine Learning to Understand & Predict HR Attrition

In today’s highly competitive business landscape, success is not just driven by innovative products or business models but also by the ability to attract and retain top talent. Talented employees are the foundation upon which every successful product is built. As such, organizations invest heavily in recruiting, training, and nurturing their workforce to maintain an edge in the market. But what happens when these valuable employees start leaving the organization? Employee attrition, if left unchecked, can be costly both in terms of resources and the loss of domain expertise.

This blog explores how machine learning can assist HR departments in predicting employee attrition, thus enabling companies to mitigate risks and make informed decisions. Specifically, we’ll take a deep dive into a case study using the IBM HR Analytics Employee Attrition and Performance dataset.

**What is Employee Attrition?**

Employee attrition refers to the gradual loss of employees over time due to various reasons like retirement, resignations, or terminations. High attrition rates are a significant concern for organizations, especially in sectors like technology, where losing a skilled employee means losing a valuable knowledge base. Attrition can be driven by multiple factors, including:

* Work-life balance issues
* Better job opportunities elsewhere
* Lack of career growth or recognition
* Unhealthy work relationships or dissatisfaction with management

By using machine learning, we can analyze employee data and predict whether someone is likely to leave the organization. This can enable HR departments to take preventive actions, improving employee retention rates.

**The Case Study: IBM HR Analytics Dataset**

In this case study, we utilize IBM’s fictional HR Analytics Employee Attrition dataset, which contains 1,470 employee records and 35 features. The primary goal of this study is twofold:

1. Identify the key factors that contribute to employee attrition.
2. Build a machine learning model that predicts employee attrition.

**Data Preparation: The Foundation of Good Analysis**

Before diving into the analytics, the dataset needs to be cleaned and prepared. The first step in this process involved checking for missing values, duplicates, and formatting issues. Fortunately, the dataset was clean, containing no missing or duplicate values. Additionally, the dataset includes both categorical and numerical features that needed to be encoded or scaled for machine learning models.

For instance, ordinal features like "Education" and "Job Satisfaction" were encoded using numerical labels, while other categorical features like "Department" were one-hot encoded.

**Exploratory Data Analysis (EDA)**

EDA helps us gain insights into the dataset before applying machine learning models. The target variable, "Attrition," was imbalanced, with 83.88% of employees staying and 16.12% leaving, making it essential to handle this imbalance during the model-building process.

Key insights from the dataset include:

* **Education and Job Role Alignment**: Employees with jobs misaligned with their educational background were more likely to leave. For example, sales representatives with unrelated educational backgrounds had higher attrition rates.
* **Age and Attrition**: The attrition rate was highest among employees aged 29 to 33, while it was lowest for those aged 34 to 45.
* **Job Role and Attrition**: Sales representatives and laboratory technicians had the highest attrition rates. Notably, 16% of research scientists left the organization, representing a significant loss of expertise.

**Machine Learning Model Development**

After cleaning and analyzing the data, we moved on to the critical step of building a machine learning model. Various classification algorithms were employed to predict employee attrition, including:

* Logistic Regression
* Support Vector Classifier (SVC)
* Decision Trees
* Random Forest Classifier
* Gradient Boosting
* AdaBoost

The dataset was split into training and test sets, and different algorithms were evaluated using accuracy, F1-score, recall, and precision. Given the imbalanced nature of the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was used to oversample the minority class (employees who left).

**Model Performance**

Among all the models, **Random Forest Classifier** emerged as the top performer with the highest F1-score and cross-validation accuracy. Below are the performance metrics for the Random Forest Classifier:

* **Accuracy**: 89.8%
* **F1-Score**: 90%
* **Precision**: 93%

Additionally, we performed hyperparameter tuning to further optimize the model, although the accuracy slightly decreased to 89.1% post-tuning. The final model was able to predict employee attrition with high precision, making it suitable for deployment in real-world HR departments.

**Feature Engineering and Dimensionality Reduction**

Several feature engineering techniques were applied to enhance model performance:

* **Outlier detection**: Z-score methods were used to detect and remove outliers.
* **Dimensionality reduction**: Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset while retaining 90% of the variance. This reduced the number of features to 21, improving model performance and reducing computation time.
* **Multicollinearity**: Variance Inflation Factor (VIF) was checked to ensure there was no multicollinearity between features.

**Key Takeaways and Business Implications**

Based on the findings from our analysis and model predictions, several actionable insights emerged:

1. **Salary Benchmark**: Setting a monthly income benchmark of $6,900 could help reduce attrition, as employees earning below this threshold were more likely to leave.
2. **Focus on Age Group 29-33**: Attrition is highest in this age group, so HR teams should focus on this segment by addressing their needs and expectations.
3. **Job Role Alignment**: Employees whose roles are misaligned with their educational background tend to have higher attrition rates. Ensuring that employees are placed in roles that align with their expertise could improve retention.
4. **Sales Department Attrition**: Almost 50% of employees in the sales department come from non-sales educational backgrounds, which could explain the higher attrition rates in this department.

**Conclusion**

HR analytics, powered by machine learning, provides organizations with the tools they need to predict and manage employee attrition. In this case study, we used Random Forest models to predict which employees were most likely to leave the organization, providing valuable insights that can help HR departments take proactive steps.

With the right machine learning model in place, companies can not only reduce employee attrition but also retain their most valuable asset—their people.

Would you like to explore more insights from your own data? Reach out, and let’s work together to build a data-driven solution tailored to your organization’s needs!