HUMAN RESOURCE ATTRITION

Created by Rajat Singh

INTRO

This project is designed to help Human Resources (HR) teams better understand and predict employee turnover, also known as employee attrition. Turnover prediction can support businesses in managing retention proactively by identifying patterns and key indicators that lead to an employee's decision to leave. This is achieved by analyzing various HR data points and building a machine learning model that can predict the likelihood of employees leaving the organization.

OUR VISION

The project uses Python and leverages powerful data-processing and machine learning libraries like Pandas, Scikit-Learn, and Matplotlib. Here's a brief overview of each:

- Pandas helps manage and manipulate the HR data, making it easy to clean, explore, and transform data points like age, job role, years at the company, and satisfaction levels.
- Scikit-Learn provides essential tools for creating, training, and evaluating the machine learning models. This library will be central to building and fine-tuning the model for accurate predictions.
- Matplotlib enables us to visualize data and model results, giving us insights into patterns like trends over time, department-level analysis, and other factors affecting employee turnover.

GOALS

To predict whether an employee will leave the company (attrition) based on various features such as age, job satisfaction, salary, etc

ABOUT DATASET

We have reduced the complexity of the dataset down to a single data file (v14). The CSV revolves around a fictitious
company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, manager name, and performance score.
Recent additions to the data include:

- Absences
- Most Recent Performance Review Date
- Employee Engagement Score

1. Data Collection and Preparation

2.Load and Explore the Data

3.Data Preprocessing

4.Split the Data into Training and Testing Sets

5. Build and Train the ModelModel

6.Feature Importance

7.Deployment

FLOW CHART

DEVELOPMENT

TRAINING PIPELINE

DATA INGESTION → MODEL EVALUATION → DATA TRANSFORMATION → MODEL TRAINING

DEPLOYMENT

Runner

GITHUB ACTIONS

GITHUB REPO

CI CD

Best model

MODEL PUSHER

MODEL EVALUATION

Best Performance

AWS EC2 → Docker image → AWS ECR → AIRFLOW → output → AWS S3

CT

DATASET OVERVIEW

- Our dataset contains information on 311 employees across 36 columns, including personal data like age and marital status, job information such as position and salary, performance metrics, and attrition data. Some columns contain missing values for employees who haven’t left yet, which we address in the preprocessing step

DATA CLEANING AND PREPROCESSING

In this stage, we start by removing irrelevant columns that don’t add predictive value, like employee names or ZIP codes. We then handle missing values, especially in the Date of Termination column

The most important being Tenure and Age. Tenure measures how long each employee has been with the company. By calculating the difference between the hire date and either the termination date or today’s date for active employees, we get a clear measure of tenure, which is often linked to turnover likelihood."

EXPLORATORY DATA ANALYSIS

Next, we explore our data visually to understand trends and relationships. First, we analyze feature distributions, like salary and tenure, to identify outliers. We also look at categorical variables, like department and performance scores, to see if turnover rates vary across these groups. Finally, we use correlation heatmaps to identify relationships between variables, for example, whether higher satisfaction scores correlate with longer tenure

MODEL BUILDING

In the model building phase, we split our dataset into training and test sets, then select features we identified as important, like salary, engagement survey scores, satisfaction, and performance. After encoding categorical features, we build several models, including logistic regression and random forests, to compare which performs best. We also tune each model's parameters to maximize accuracy and precision.

MODEL EVALUATION

After training, we evaluate each model's performance using metrics like accuracy, precision, recall, and AUC-ROC to understand its effectiveness. For example, recall is critical here, as we don't want to overlook employees likely to leave. We also analyze feature importance to understand which factors impact turnover most, which will be helpful for HR's intervention efforts.

DEPLOYMENT - MLOPS PIPELINE WITH AWS AND DOCKER

To deploy this project, we implemented a robust MLOps pipeline on AWS, which enables continuous integration and deployment. Here's a step-by-step overview:

1. EC2 and Git Runner Workflow: The codebase is hosted on GitHub, and we set up a GitHub Runner workflow to automatically deploy the project on an AWS EC2 instance. Every new update in the repository triggers this workflow, keeping the deployment current.
2. Containerization with Docker: To ensure consistency and portability, the entire project runs in a Docker container. Docker packages the app with all dependencies, so we can deploy it seamlessly across different environments.
3. Training with Apache Airflow: We used Airflow to orchestrate the training pipeline. Airflow schedules, triggers, and monitors each step, from data preprocessing to model training. This approach allows for flexible retraining of the model on updated data.
4. Model and Prediction Storage with S3: After training, the best model is automatically saved in an S3 bucket, along with predictions and other outputs. This cloud storage approach ensures secure, accessible storage of model artifacts and makes it easy to retrieve the latest model for future use

KEY INSIGHTS

Some important insights we found include:

1. Salary and satisfaction scores significantly impact turnover risk.
2. Employees with low engagement scores or high absences are more likely to leave.
3. Certain departments have higher turnover, potentially highlighting high-stress roles.

These findings help HR prioritize areas where employee retention can be improved."

CONCLUTION

In conclusion, this project demonstrates how machine learning can empower HR teams with actionable insights on employee turnover. From data processing to model evaluation and deployment, each step in this pipeline contributes to creating a reliable, predictive model. This tool can be a valuable addition to HR’s toolkit, enabling proactive retention strategies."

THANK YOU!

FOR YOUR ATTENTION