Predicting Employee Attrition

Submitted By:
Anushka Saraswat
(saraswatanushka007@gmail.com)

Video Link : 🗎 20240511_194332.mp4

ABSTRACT

Problem: Create a model for predicting Employee Attrition

Dataset Used: IBM HR Analytics Employee Attrition & Performance
Source
(https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?resource=download)

Model: Random Classifier Model used

Key Findings:

- The columns ‘Over18’, ‘EmployeeCount’, ‘EmployeeNumber’, ‘StandardHours’ don’t have any role in predicting employee attrition.
- Most of the employees that are leaving the company belong to the age group of 28- 35
- Eployees from the Research and Development department are leaving the department as compared to other departments.
- There is no gender discrimination against female employees, in fact the percentage of male employees leaving the company is slightly more than the female employees.
- Dataset needs to be resampled as it is highly imbalance against the attrition target variable.

Introduction

Problem: Employee attrition, or turnover, is a significant concern for many organizations. It can lead to a number of negative consequences, including:

- **Increased costs:** The cost of recruiting, hiring, and onboarding new employees can be substantial.
- **Decreased productivity:** When employees leave, it can take time for new hires to reach full productivity. This can lead to a temporary dip in overall team or company output.
- **Loss of valuable knowledge and experience:** Departing employees take their knowledge and experience with them. This can create a knowledge gap that can be difficult to fill.

Context and Background: In today's competitive job market, retaining top talent is crucial for organizational success. By being able to predict which employees are at risk of leaving, companies can take proactive steps to retain them. This might involve offering competitive salaries and benefits, providing opportunities for professional development, or creating a more positive work environment.

Objectives: This project aims to develop a machine learning model to predict employee attrition. Here are the specific objectives:

- **Analyze employee data to identify factors that contribute to employee attrition.** This might include factors such as job satisfaction, salary, workload, and opportunities for advancement.
- **Develop a machine learning model to predict which employees are likely to leave the company.** This model can be used to identify employees who are at risk of leaving the company

Data Analysis

The dataset that is used is *'IBM HR Analytics Employee Attrition & Performance'* source for this dataset is *'https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?resource=download'*. Size of the dataset is 51kB.

Preprocessing Steps Undertaken :

- Checked for null and missing values, no such values found
- Performed label encoding on categorical values
- Removed the unnecessary columns

Key Findings From Data Exploration:

- The columns 'Over18', 'EmployeeCount', 'EmployeeNumber', 'StandardHours' don't have any role in predicting employee attrition.
- The columns 'StandardHours', 'Over18', 'EmployeeCount' have got the same value repeating
- Most of the employees that are leaving the company belong to the age group of 28- 35
- Eployees from the Research and Development department are leaving the department as compared to other departments.
- There is no gender discrimination against female employees, in fact the percentage of male employees leaving the company is slightly more than the female employees.
- Dataset needs to be resampled as it is highly imbalance against the attrition target variable.

Trends in the Dataset

- Employees with high Job Level have high monthly income
- Employees with high Perfromance Rating have high Perfromance Salary Hike
- Employees with more total working hours have a higher job level and monthly income

Model Develpment

Choice of model : Random Forest Classifier

Reason : Well Random Forest Classifier is a good model for working with data having numerical and categorical values. Also it prevents overfitting and does not use a single feature as the dominant one but choose them randomly for different trees. It provided a good accuracy, precision and recall.

Training Process :

- First of all I have splitted the dataset into X and Y, Y having the target variable and X having the rest of the columns.
- ThenI I have done oversampling of the dataset using RandomOverSampler from imblearn.
- Oversampling is done because the dataset was small in size and very imbalance with respect to the target variable.
- Then I have again splitted the dataset for training and testing. 75% for training and 25% for testing.
- Then I have fitted the training part in the model and checked the accuracy, it came out to be 99% something
- Then I have checked the accuracy with testing part and it also gave a good accuracy of 96.5%
- Then I have tried removing some features from the dataset and training the model again, although these features didnt have a huge impact on the working of model but still they had some minute impact.

Model Evaluation:

- Training dataset accuracy : 0.9989183342347214
- Testing dataset accuracy : 0.965964343598055
- Confurion Matrix:

	Actual Positive	**Actual Negative**
Predicted Positive	293	18
Predicted Negatice	8	298

- Classification Report :

	Precision	**Recall**	**F1- score**	**Support**
0	0.97	0.94	0.95	311
1	0.94	0.97	0.95	306
Accuracy			0.95	617
Macro avg	0.95	0.95	0.95	617
Weighted avg	0.95	0.95	0.95	617

Discussion and Insights

- Employees from the age group of 28-35 are more prone to leaving the company as compared to the older and younger ones.
- People who are in the company for 5 years are less likely to leave
- Employees with more working years have a good job level and monthly income
- Employees at a good job level are less likely to leave