Project 1: Proposal & Data Selection

Kesav Adithya Venkidusamy

Bellevue University - Master of Science in Data Science

DSC680-T301 Applied Data Science (2231-1)

Professor Catherine Williams

09/04/2022

Topic

IBM HR Analytics and Prediction of Employee Attrition

Business Problem

Employee retention strategies are integral to the success and well-being of a company. There are often many reasons why employees leave an organization, and in this case study, I will explore some of the key drivers of employee attrition. Employee attrition measures how many workers have left an organization and is a common metric companies use to assess their performance. Some of the common reasons for employee attrition are as follows.

1. Poor job satisfaction and pay
2. Not enough career opportunity
3. Poor workplace culture
4. Lack of employee motivation
5. Poor work-life balance
6. Not fitting in and feeling sense of belonging

There are three main types of employee attrition.

1. **Involuntary attrition:** Involuntary attrition happens when the company decides to part ways with the employee. Rather than the employee deciding to leave, it is the company's decision to let go of the employee. This can be due to position elimination, termination or layoffs.
2. **Voluntary attrition:** Voluntary attrition happens when an employee decides to leave the company. This can be for many reasons like accepting a new job offer, making a career change, relocation.
3. **Retirement attrition:** Retirement attrition happens when employees reach their stage in life for retirement.

While turnover rates vary from industry to industry, the Bureau of Labor Statistics reported that among voluntary separations the overall turnover rate was 32.7% in 2021, and even more than this in

2022. So, predictive attrition model helps in not only taking preventive measures but also into making better hiring decisions. Minimizing attrition can ensure associates stay longer, enabling them to continue benefiting the organization operations.

Datasets

The dataset is extracted from the following Kaggle website. This is a fictional dataset created by IBM data scientists. The dataset contains approximately 1500 entries.

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

This dataset presents an employee survey from IBM, indicating if there is attrition or not. Using this dataset, I will uncover the factors that lead to employee attrition and explore some of factors contribute to the attritions.

Characteristics

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer
Associated Tasks	Classification
Number of Instances	1470
Number of Attributes	35
Missing Values	No
Area	Social

Attributes information

Feature Name	Feature Description	Feature Type
Age	Age of the person	Continuous
Attrition	Person has left the company or not	Target
BusinessTravel	How frequently the person travels	Discrete
DailyRate	Daily Rate for the employee	Continuous
Department	Department of the person	Discrete
DistanceFromHome	Distance of the company from home	Continuous
Education	Education of the person	Discrete

	1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'	
EducationField	Education field of the person	Discrete
EmployeeCount	Count of Employee	Discrete
EmployeeNumber	Employee id of the person	Continuous
EnvironmentSatisfaction	Environment Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Discrete
Gender	Gender	Discrete
HourlyRate	Hourly rate for the employee	Continuous
JobInvolvement	Involvement in job 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Discrete
JobLevel	Job Level	Discrete
JobRole	Job Role	Discrete
JobSatisfaction	Job Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Discrete
MaritalStatus	Marital Status of Employee	Discrete
MonthlyIncome	Monthly Income of the person	Continuous
MonthlyRate	Monthly Rate	Continuous
NumCompaniesWorked	Number of Companies worked	Discrete
Over18	Over 18 years	Discrete
OverTime	Worked over time	Discrete
PercentSalaryHike	Percentage of Salary Hike	Continuous
PerformanceRating	Performance Rating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'	Discrete
RelationshipSatisfaction	Relationship Satisfaction for the employee 1 'Low' 2 'Medium' 3 'High' 4 'Very High	Discrete
StandardHours	Standard work hours	Discrete
StockOptionLevel	Stock Option Level given to the employee	Discrete

TotalWorkingYears	Total Number of years worked	Continuous
TrainingTimesLastYear	Training times attended during last year	Continuous
WorkLifeBalance	Work Life Balance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'	Discrete
YearsAtCompany	Number of years with current company	Continuous
YearsInCurrentRole	Number of years in the current role	Continuous
YearsSinceLastPromotion	Number of years since last promotion	Continuous
YearsWithCurrManager	Number of years with current manager	Continuous

Methods

Following modelling techniques will be used on the dataset to determine which features are mostly related or correlated to our target variable “Attrition”.

1. Logistic Regression
2. Decision Tree
3. Random Forest

Logistic regression is a statistical analysis method used to predict a binary outcome such as yes or no based on prior observation of the data set. Here, “Attrition” feature present in the dataset has only binary values: whether the person has left the organization or not. So, this feature will be used as target for the model. This model falls under supervised learning as the data is well labelled and has a target variable, a column in the data representing values to predict from other columns in the data. Under supervised learning, this dataset falls under classification model as it reads the input and generates an output that classifies the input into two categories: one having attrition as “Yes” and another as “No”.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned

In addition to running the above models with 1 target and 34 features, evaluation will be performed by executing these models with only 5 best features where the 5 features are selected based on highest chi-squared statistics.

Ethical Considerations

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions make from this project's study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project to larger audience, the underlying assumptions and data considerations should be shared.

Another ethical consideration when dealing with employee information is to ensure no personal and sensitive information is present in the dataset. Since this dataset is fictional created by IBM data scientists, this is already taken care by them and personal identifying information (like gender, age) is broad enough which is untraceable to any individual.

Challenges/Issues

One of the earliest challenges we might face is during the data preparation step of the model building. Identifying the correct features that contribute to the target, planning on how to handle the insufficient data, deciding the next steps if the data is imbalanced to name a few. The way to mitigate these issues would be creating various visualizations to identify correlations. To mitigate data imbalance, we may choose to over-sample or under-sample the dataset. We may also need to go back to research other relevant supplement datasets to strengthen the cause.

Reference

Dataset: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Types and reasons for Attrition: https://www.betterup.com/blog/employee-attrition

Random Forest: https://en.wikipedia.org/wiki/Random_forest

Bureau of labor statistics: https://www.bls.gov/news.release/jolts.t18.htm