HR Attrition Analysis

Domain Background

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. If an employee leaves a company in this scenario, it is a waste of investment in terms of both time and money for that company. What if there's a way to prevent this scenario from happening at all? What if a company is able to predict if an employee is likely to leave it even before he/she decides to do it himself/herself? This is where we utilize the machine learning algorithms to help us achieve our objective.

Personally, I chose this problem because of my active involvement in this domain for the past year. Like in most of the organizations, this is an active field of research in my organization as well and I'm an integral part of the team working on this domain and it's imperative for us to try and find some sort of a solution to this problem.

Problem Statement

The company wants to understand what factors contributes most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees which can be achieved if one can somehow predicts those exact factors which may lead to an employee leaving the company. Overall, the implementation of this model will allow management to create better decision-making actions.

Datasets and Inputs

(Data source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.)

IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. To use a matrix structure, we changed the model to reflect the following data:

Name	Description
AGE	Numerical Value
ATTRITION	Employee leaving the company (0=no, 1=yes)
BUSINESS TRAVEL	(1=No Travel, 2=Travel Frequently, 3=Tavel Rarely)
DAILY RATE	Numerical Value - Salary Level
DEPARTMENT	(1=HR, 2=R&D, 3=Sales)
DISTANCE FROM HOME	Numerical Value - THE DISTANCE FROM WORK TO HOME

EDUCATION	Numerical Value
EDUCATION FIELD	(1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL)
EMPLOYEE COUNT	Numerical Value
EMPLOYEE NUMBER	Numerical Value - EMPLOYEE ID
ENVIROMENT SATISFACTION	Numerical Value - SATISFACTION WITH THE ENVIROMENT
GENDER	(1=FEMALE, 2=MALE)
HOURLY RATE	Numerical Value - HOURLY SALARY
JOB INVOLVEMENT	Numerical Value - JOB INVOLVEMENT
JOB LEVEL	Numerical Value - LEVEL OF JOB
JOB ROLE	(1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE)
JOB SATISFACTION	Numerical Value - SATISFACTION WITH THE JOB
MARITAL STATUS	(1=DIVORCED, 2=MARRIED, 3=SINGLE)
MONTHLY INCOME	Numerical Value - MONTHLY SALARY
MONTHY RATE	Numerical Value - MONTHY RATE
NUMCOMPANIES WORKED	Numerical Value - NO. OF COMPANIES WORKED AT
OVER 18	(1=YES, 2=NO)
OVERTIME	(1=NO, 2=YES)
PERCENT SALARY HIKE	Numerical Value - PERCENTAGE INCREASE IN SALARY
PERFORMANCE RATING	Numerical Value - ERFORMANCE RATING
RELATIONS SATISFACTION	Numerical Value - RELATIONS SATISFACTION
STANDARD HOURS	Numerical Value - STANDARD HOURS
STOCK OPTIONS LEVEL	Numerical Value - STOCK OPTIONS
TOTAL WORKING YEARS	Numerical Value - TOTAL YEARS WORKED
TRAINING TIMES LAST YEAR	Numerical Value - HOURS SPENT TRAINING
WORK LIFE BALANCE	Numerical Value - TIME SPENT BEWTWEEN WORK AND OUTSIDE
YEARS AT COMPANY	Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPNAY
YEARS IN CURRENT ROLE	Numerical Value -YEARS IN CURRENT ROLE
YEARS SINCE LAST PROMOTION	Numerical Value - LAST PROMOTION
YEARS WITH CURRENT MANAGER	Numerical Value - YEARS SPENT WITH CURRENT MANAGER

Solution Statement

I plan to run a series of classification models to determine the probability of a certain employee to fall into the condition of Attrition and thus its high risk of leaving the company. We will then test different parameters and probability threshold using confusion Matrixes, Area under the Curve and F-score/Accuracy to determine which of these models is the best predictor and will recommend its use in practice. I'm also planning to extract the topmost important features which contributes towards the attrition for a given employee. Please find below some of the models that I'm going to run on the given input data:

- Logistic Regression
- Random Forest

- AdaBoost
- XGBoost

Benchmark Model

The chosen problem is available as a competition on Kaggle and I have chosen one of the student's public submission as a tool for benchmarking my own solution. Please find below the link to that particular solution:

https://www.kaggle.com/nitya123/ibm-hr-analytics-employee-attrition-performance#Business-Problem

The user ran three classification models on the given problem and below are the stats for each of them:

Logistic AUC: 0.58

Decision Tree AUC: 0.57

Random Forest AUC: 0.57

The user has also tried to create the feature importance maps for the above models. I'll take this numbers as a benchmark and will try to improve on these numbers and also create my own feature importance graphs to see if I get any different or better results.

Evaluation Metrics

This is a binary classification problem and hence I have opted for some of the evaluation metrics suitable for classification problems. Please find details of some of them that I'm going to use to evaluate my models:

- ROC-AUC: Area under the curve for ROC (Receiver Operating Characteristics) graph. ROC can be plotted using TPR and FPR as x and y axis of a graph. Please find below their calculation formulas:
    - TPR (True Positive Rate): TP/ TP + FN
    - FPR (False Positive Rate): FP/ TN + FP
        - TN – True Negative, FP – False Positive, TP – True Positive, FN – False Negative
- Accuracy score: Denotes the accuracy of the model and can be calculated as below:
    - (TP + TN) / (TP + TN + FP + FN)
- Sensitivity (True Positive Rate): Sensitivity in this case can be quite helpful as one might want to know the employee quit accuracy which is what we originally wanted to predict.

- TP / (TP + FN)

Project Design

Please find below the detailed outline of my design approach:

- **Data Analysis:**
    - Finding out what types of variables there are in this dataset and divide them into categorical and numerical variables.
    - Further try to subclassify numerical variables into discrete and continuous.
    - Find out missing values in the data if any, by checking for any null/blank values present inside the data
    - Try to find outliers present in the data by plotting boxplots.
- **Splitting up data:**
    - Split up the data into the test and training sets, I'll mostly go with the 80-20 split with 80% of the data being used for training the model.
- **Data Cleanup:**
    - Handle any missing values found during the analysis step both for test and training datasets.
        - For numerical values, depending on how much data is missing or null, I'll go with either random sample imputation (< 50%) or impute missing values by value far in the distribution (> 50%).
        - Similarly, for categorical data, I'll go for the most frequent category to replace the missing values in case the missing values are less, or else I'll create a new category called 'Missing' if we have a lot of missing values in a field.
    - Handling outliers:
        - For numerical variables, if the field contains outliers and its distribution is Gaussian, I'll go for top coding technique according to which I'll replace the outlier values with the corresponding maximum/minimum value obtained from boxplot range.
        - On the other case, if a variable containing an outlier seems to have a skewed distribution, I'll try to discretize the data into separate quantiles to make the distribution more normal and try to adjust those outlier values to their closest quantile range.
        - For categorical variables, depending on whether an independent variable contains a lot of rare labels or only a few, I'll either replace those labels by the most frequent category or create a new label called 'rare' to denote those infrequent categories.
- **Data Encoding:**

  - Once the cleanup part is done, it is time to encode the categorical variables into numerical values as our machine learning models operate better on numeric values.
  - Depending on the number of categories present in an input field, I'll use the below two techniques for encoding.
    - One-Hot encoding: If the number of categories is less, then I'll use this technique which will simply create a new variable for each of the category and replace the values by 0 or 1 depending on that category was present or not in that particular row.
    - Replacement by risk probability: If the number of categories is huge, then one-hot encoding might lead to an explosion of fields and can increase the model complexity, so instead, I'll replace the individual values with the mean distribution of each category with respect to the output variable.
- **Feature Scaling:**
  - If required, I'll use the min-max scaler to scale the numerical values between the range 0 –1. This is really helpful for linear algorithms like Logisitic Regression but may not be required for decision-tree based algorithms.
- **Model Creation and Evaluation:**
  - As described above, I'll try to build the following models once the data preprocessing part is done:
    - Logistic Regression
    - Random Forests
    - Adaboost
    - XGBoost
  - I'll evaluate my models using the below metrics:
    - ROC-AUC
    - Accuracy
    - Sesitivity
  - I'll also create the feature importance graph to visualize the top most important features.
  - Finally, I'll try to get the contributing factors which led to the final model decision for each line item.

Additional Citations

https://towardsdatascience.com/people-analytics-with-attrition-predictions-12adcce9573f

http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html