Employee Attrition

Predictive Analysis

Sowmya Chavali
DSC-680
Bellevue university

Overview

Hiring and retaining top talent is an extremely challenge task for the Human Resources department, which requires capital, time, and skills. Business owners spend up to 40% of their time on simple tasks not resulting in direct profit, such as hiring. For a new hire 15-20% of their salary is spent on just recruitment. For a better performance of the companies the HR department has to take measures to reduce employee attrition.

Introduction

It is believed that most companies would like to understand why their best employees voluntarily leave. This is one of the widely used case studies to find why it is happening and which factors are being responsible for that. I have found various studies and articles on the internet to solve this problem finding factors responsible for that. According to the article by Balance Articles, employees jump across jobs and roles for averagely 12 times during their lifetime career. It clears that leaving and finding new jobs wasting billions of dollars. Therefore, managing their employee attrition is very important in the company.

This HR case study is created to leverage the power of data science to reduce employee turnover and optimize the HR department. Idea behind this case study is to develop a model which predicts employees that are more likely to quit.

Problem Statement

In this study, we attempt to understand the likelihood of an employee leaving the company and identifying the key indicators for the attrition. Post this analysis the company can adopt strategies to improve on employee retention.

This is a supervised classification problem where the target is a binary variable, 0 and 1 for active employee and ex-employee. In this study, our target variable is the probability of an employee leaving the company.

Data Understanding

In this case study, the IBM HR dataset was sourced from Kaggle[1].

Dataset contains employee data for 1,470 employees with various information about the employees. I will use this dataset to predict when employees are going to quit by understanding the main drivers of employee churn.

As stated on the IBM website[2] "This is a fictional data set created by IBM data scientists". Its main purpose was to demonstrate the IBM Watson Analytics tool for employee attrition.

The dataset contains 1,470 rows and 35 columns.

Columns

Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager.

Data Preprocessing

It is important to process the data so as it to make it fit for use. For the HR Analytics dataset needs some cleaning and renaming so as to make it fit for the data mining algorithms. The data pre-processing steps I applied are-

→ Checking for Nulls and outliers

→ Changing categorical values to numeric

→ Identifying and removing columns with single values

Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Correlation between features is identified to understand what factors has impact on the target. Here we remove the irrelevant or less important features which do not contribute much to the target variable. This step is crucial to achiev better model accuracy.

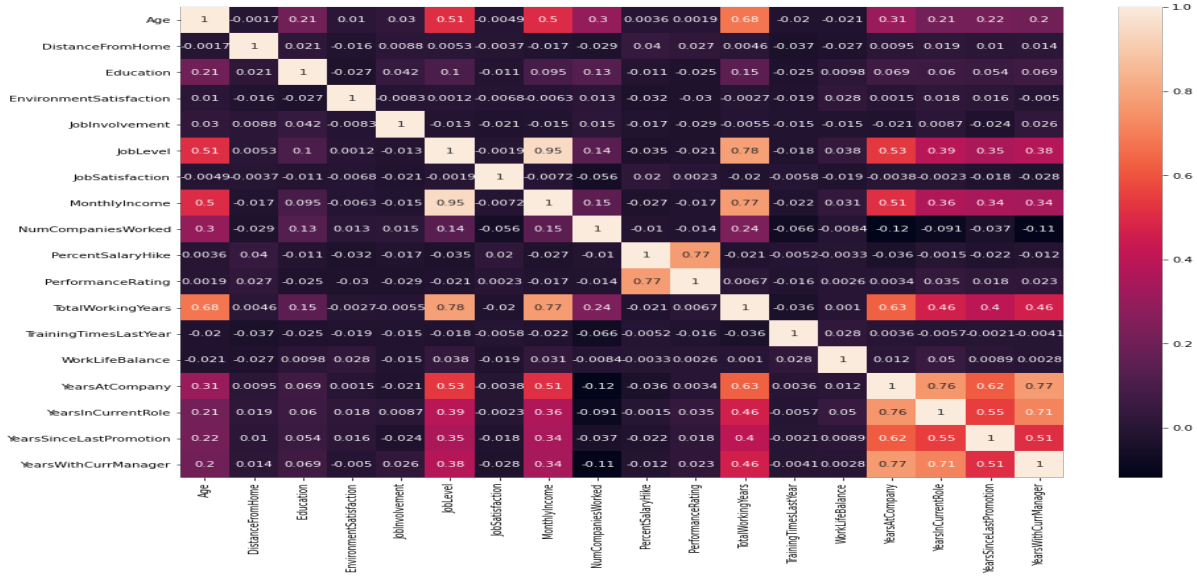
Feature Importance -

Monthly Income	0.101
Age	0.082
Over Time	0.071
Total Working Years	0.069
Distance From Home	0.068
Years At Company	0.057
Percent Salary Hike	0.051
Environment Satisfaction	0.043
Job Satisfaction	0.04
Years Since Last Promotion	0.04
Years With Curr Manager	0.04
Work Life Balance	0.037
Years In Current Role	0.037
Job Involvement	0.036
Education	0.031
Job Level	0.027
Performance Rating	0.008

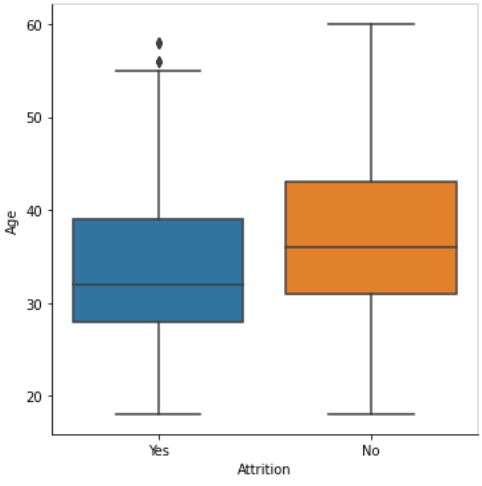
Data analysis

Before identifying the key indicators for attrition, I performed EDA to get more insight of the data. Therefore, a data analysis is done to get more acquainted and familiar with the reduced and selected data.

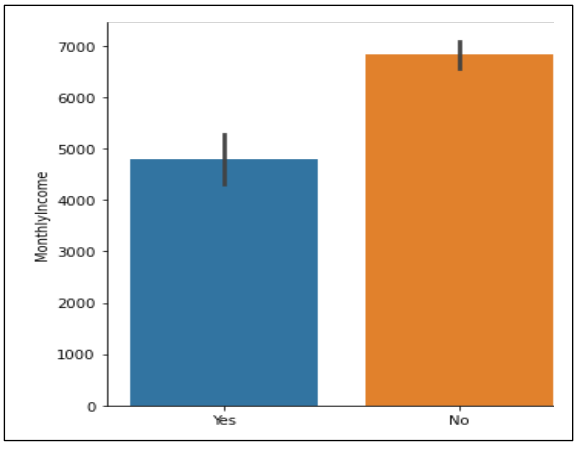
Heatmap



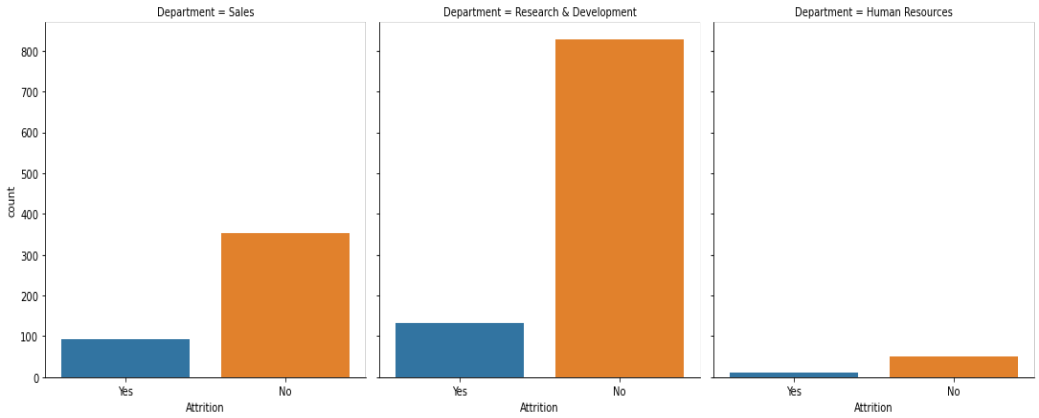
Age vs attrition



Income vs attrition



Attrition by department



Data Modelling

Modeling is one of the best ways to validate our work. For this, we need to split the datasets into two parts called training set (75%) and testing set (25%). Basically, a training set is used to train the model, and the test set is used to test the model. This is to avoid the overfitting the model to reach high accuracy. After completing data wrangling we split the sample data following ways.

Splitting the train and test sets

X_train, X_test, y_train,y_test= cross_validation.train_test_split(X,y,test_size=0.25)

Once we have done the split, we will proceed to train and validate the model.

Metrics

Many metrics are used to measure the accuracy of the model, that is measuring the predictive accuracy of the model.

Metrics that are used to test the models:

1. Sensitivity / Recall
2. Specificity
3. Precision
4. F1 Score

It is important to review these metrics to decide if the model is performing well.

We used several models to achieve our goal. The models that are used in this analysis are:

a) Logistic Regression

Logistic regression is a simple and more efficient method for binary and linear classification problems.

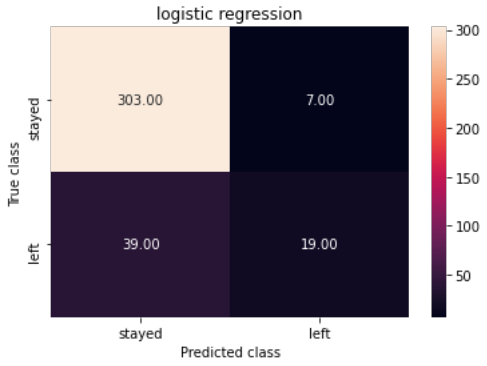
With the Logistic regression we achieved an accuracy of 87.5%.

Classification report

	precision	recall	f1-score	support
No	0.89	0.98	0.93	310
Yes	0.73	0.33	0.45	58
accuracy			0.88	368
macro avg	0.81	0.65	0.69	368
weighted avg	0.86	0.88	0.85	368

We can see that the model predicts quite well the employees that won't quit(93% accuracy) but not the other way round i.e., for the employees leaving the company the accuracy is only 45%.

Confusion Matrix



b)Random Forest

Random forest aggregates Classification (or Regression) Trees. A decision tree is composed of a series of decisions that can be used to classify an observation in a dataset.

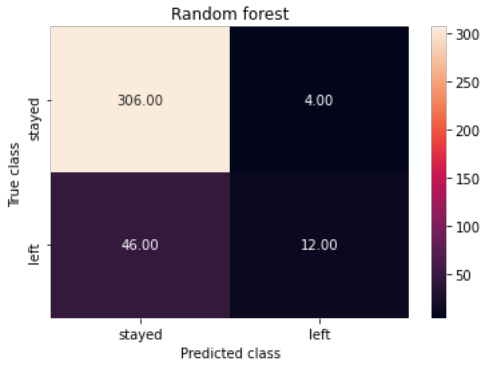
With Random forest we achieved an accuracy of 86.4%.

Classification report

	precision	recall	f1-score	support
No	0.87	0.99	**0.92**	310
Yes	0.75	0.21	**0.32**	58
accuracy			0.86	368
macro avg	0.81	0.60	0.62	368
weighted avg	0.85	0.86	0.83	368

Again, the model only predicts quite well the employees who stay with the organization which is at 92%. But fails predicting the other way round(32%).

Confusion matrix



c)Ada boosting

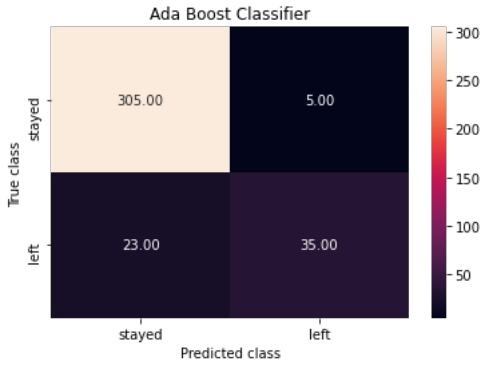
The method automatically adjusts its parameters to the data based on the actual performance in the current iteration. Meaning, both the weights for re-weighting the data and the weights for the final aggregation are re-computed iteratively.

With Ada boosting we achieved an accuracy of 92.3%.

Classification report

	precision	recall	f1-score	support
No	0.93	0.98	0.96	310
Yes	**0.88**	**0.60**	**0.71**	58
accuracy			0.92	368
macro avg	0.90	0.79	0.84	368
weighted avg	0.92	0.92	0.92	368

Confusion Matrix



This is a decent solution. Using the Ada Boost lassifier, the performance has been improved. Boosting has allowed us increase the accuracy and also the recall.

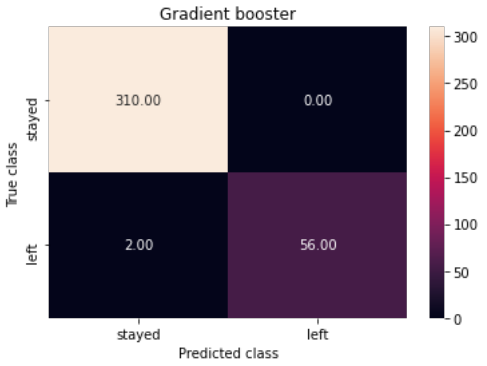
Gradient Boosting

The technique yields a direct interpretation of boosting methods from the perspective of numerical optimization in a function space and generalizes them by allowing optimization of an arbitrary loss function.

With Gradient boosting we achieved an accuracy of 99 %.

	precision	recall	f1-score	support
No	0.99	1.00	1.00	310
Yes	**1.00**	**0.97**	**0.98**	58
accuracy			0.99	368
macro avg	1.00	0.98	0.99	368
weighted avg	0.99	0.99	0.99	368

Confusion Matrix



Gradient booster classifier is the chosen model as the model accuracy increased to 99% also recall.

Results

Predictions are made on employees who are more likely to leave or stay in the company. As the company generates more data on its employees, the algorithm can be re-trained using the additional data and generate more accurate predictions.

- Monthly income, age and Over time are the primary reasons for attrition.
- With Logistic regression model we correctly predict the employee leaving the company or not with 87.5 % accurate results.
- With Random forest model we correctly predict the employee leaving the company or not with 86.4 % accurate results.
- With Ada boosting model we correctly predict the employee leaving the company or not with 92.3 % accurate results.
- With Gradient boosting model we correctly predict the employee leaving the company or not with 99 % accurate results.

References

1. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv
2. https://community.ibm.com/community/user/home
3. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/feature-selection-data-mining
4. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
5. https://analyticsindiamag.com/adaboost-vs-gradient-boosting-a-comparison-of-leading-boosting-algorithms/