Employee Attrition Prediction

Santosh Omprakash

DSC 680 – Applied Data Science

Bellevue University

Introduction

Employee attrition is the departure of employees from organization over time. Employee attrition can be voluntary or involuntary. Involuntary attrition is result of organization's action where the employee is let go due to various reasons like performance. Voluntary attrition is when employee leaves the organization based on their discretion. Employee attrition that is voluntary can be due to various reasons like personal, job mismatch, retirement. Employee attrition is common in all organizations and if it is not managed properly, it can result in key members leaving the organization which in turn would result in decreased productivity. The organization then have to hire and train new people which results in increased cost and is time consuming process.

The project proposal is to predict employee attrition based on which companies can take certain actions to reduce attrition. As part of this project, company's HR dataset will be analyzed to predict if an employee is likely to leave the organization. The dataset is first preprocessed to remove unnecessary variables. Data visualization is performed using Exploratory Data Analysis (EDA). There are many features in the dataset, and they will be analyzed using EDA to look at some factors that affect attrition. The features required for machine learning model is selected and dataset is split into train and test. Machine learning models like Random Forest, Linear Regression will be applied, and performance will be evaluated. The model with best score is selected and tuned to improve performance. The tuned model is evaluated for prediction results.

Problem Statement

Employee attrition is the decrease in size of workforce. Employee can leave the organization due to various reasons. There are many factors like age, salary, workplace challenges, work life

balance, etc., which lead to whether an employee decides to stay with organization or not. Human resource is the important factor and considerable time is spent to recruit employees.

As per the reports, the national average attrition rate in US was 27% in 2018 and 36% in 2019. Employee attrition is increasing over the years and a survey in US indicates that in 2030, there could be $430 billion loss annually due to low talent retention. Organizations always focus to have their professional employees stay to maintain productivity and to reduce cost involved in hiring and training new people. There can be impact on project schedule due to recruiting and training of new employees.

There can be different human resource problems which cannot be determined using specific scientific formula. Hence, machine learning is the best way to solve this problem. The 'IBM HR' data from Kaggle is used for analysis to predict employee attrition. Machine learning algorithms are used, scores of few models are compared and best model is selected.

Data Exploration

The 'IBM HR Analytics Employee Attrition & Performance' dataset is selected from Kaggle. The dataset is explored to look at different variables. The dataset has 35 variables, they are –

Age, 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, YearsWithCurrentManager.

Attrition – is the predictor variable that contains value Yes and No. ‘Yes’ indicates employee attrition and ‘No’ indicates employee staying with the company.

The dataset has 1,470 records. It was first explored to check for null values. There are no null values in any of the variables. Some of the variables were removed from further processing as it does not add much value in attrition prediction. The features removed are - EmployeeCount, Over18, StandardHours, EmployeeNumber.

Exploratory Data Analysis (EDA)

The predictor variable ‘Attrition is visualized to check for records.

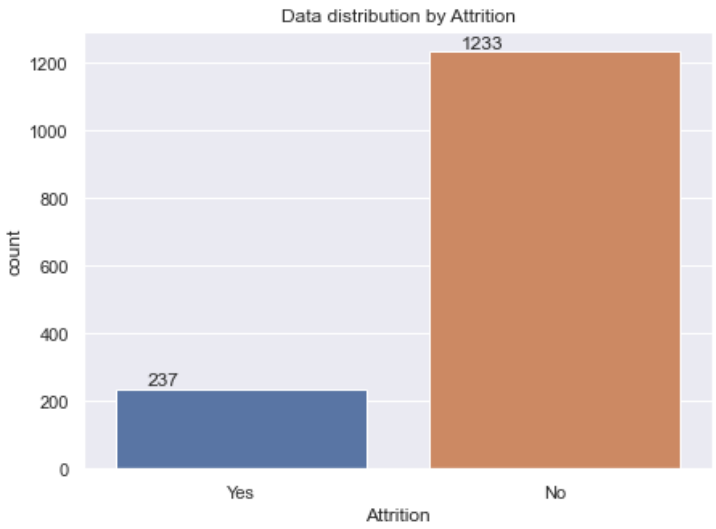

Figure 1

As shown in figure 1, dataset has 237 Attrition records and 1233 records non-attrition records. Hence, the dataset is highly imbalanced.

Data visualization of OverTime to Attrition is plotted.

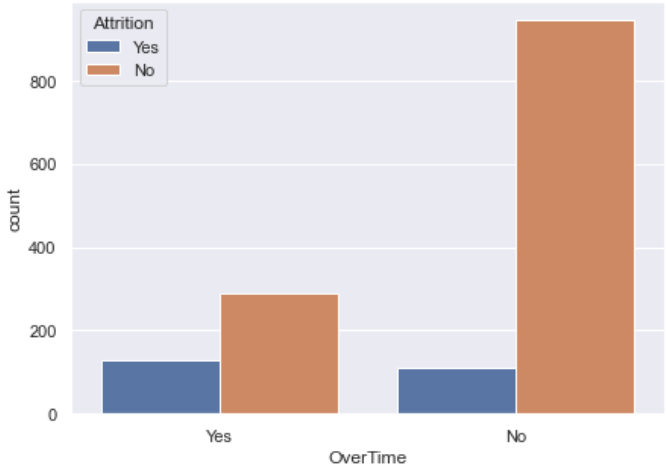

Figure 2

As shown in figure 2, Attrition due to over time is more.

YearSinceLastPromotion variable is then visualized. As shown in figure 3, Employees who are recently promoted are likely to stay in the company.

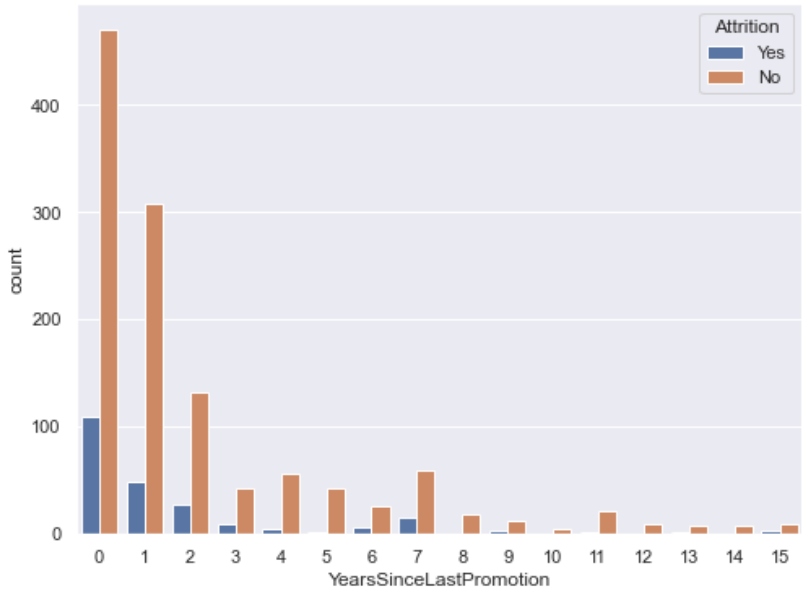

Figure 3

Age feature is visualized as shown in figure 4.

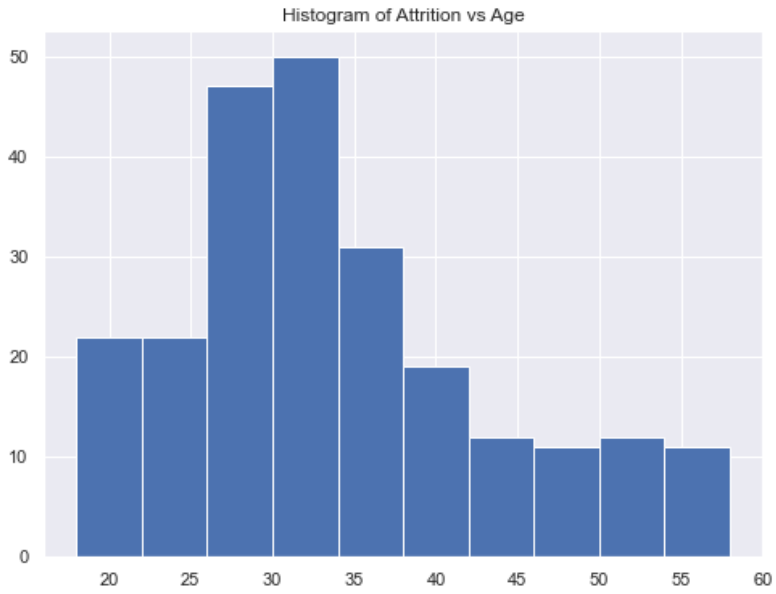

Figure 4

As seen, attrition is more between age group 25 to 35.

Density plot of Attrition with regards to MonthlyIncome is shown in figure 5. It shows that attrition is more for monthly income between $2000 to $4000.

Figure 5

Figure 6 shows that attrition is more when number of companies worked is less. Hence, employee is most likely to leave organization during early stages in their career.

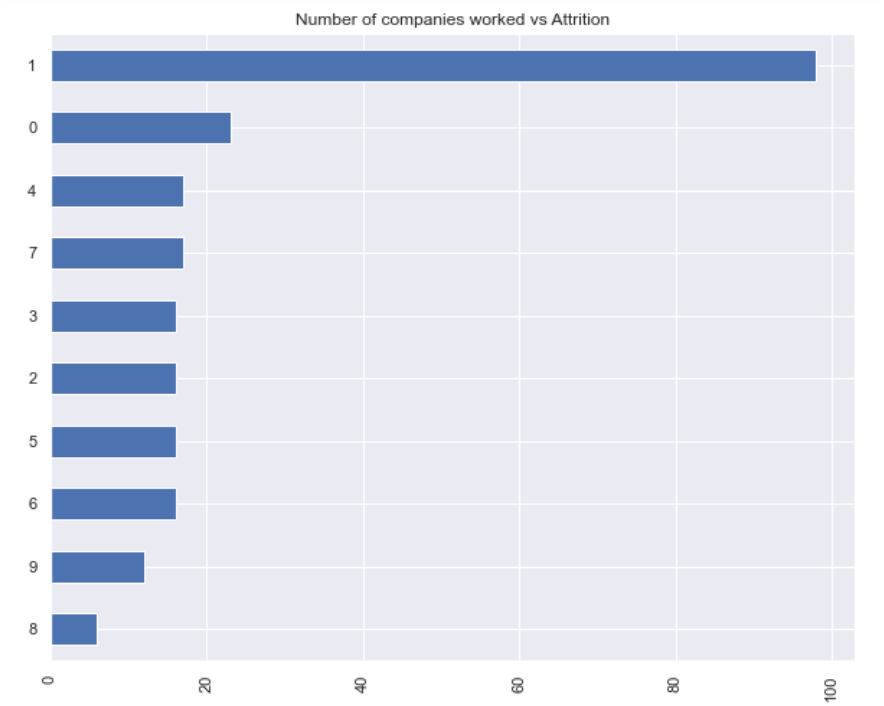

Figure 6

Attrition related to EducationField is shown in figure 7. It shows that attrition is more in Life Sciences and Medical. It is least in Human Resources.

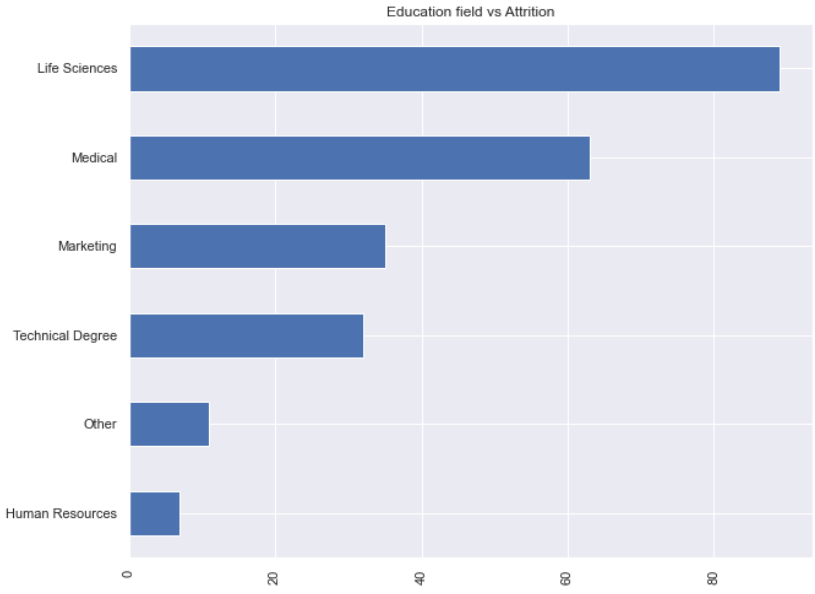

Figure 7

Feature selection is performed using SelectKBest technique. For this process, the best 20 features from the dataset are selected. Figure 8 shows the features selected for processing.

	Feature_Name	Score
18	OverTime	56.559538
23	TotalWorkingYears	42.143368
0	Age	41.729236
29	YearsWithCurrManager	34.427882
27	YearsInCurrentRole	33.692718
11	JobLevel	33.368331
15	MonthlyIncome	31.017502
26	YearsAtCompany	28.952703
14	MaritalStatus	28.049293
22	StockOptionLevel	24.439454
10	JobInvolvement	18.656350
7	EnvironmentSatisfaction	15.169368
13	JobSatisfaction	12.686817
4	DistanceFromHome	8.900725
25	WorkLifeBalance	7.622819
2	DailyRate	6.015380
3	Department	5.054415
17	NumCompaniesWorked	2.784982
12	JobRole	2.690796
21	RelationshipSatisfaction	2.212336

Figure 8

Model Selection and Evaluation

The variables with text data are converted to numerical values using Label Encoder. The dataset is split into train and test. 80% of the data was selected for training and 20% was selected for testing. Stratify option is used when splitting train and test data so that the proportion example is present in train and test data. Since the dataset variables have different range of values, it is important to get them to common value range. Hence, normalization is performed using StandardScaler. The metric used for machine learning model is ROC AUC (receiver operating characteristic curve). The train data is fit with different machine learning models.

The models used are Random Forest, Decision Tree Classifier, SGD classifier and Logistic Regression.

As show in figure 9, Logistic Regression model has the best ROC-AUC score of 80%.

	roc_auc
RandomForestClassifier	0.792508
DecisionTreeClassifier	0.624336
SGDClassifier	0.732926
LogisticRegression	0.804756

Figure 9

Hence, Logistic Regression is selected for further performance tuning. Logistic Regression model is tuned with hyper parameters using Grid Search CV. The parameters selected are C = 0.1, penalty = l2 and solver = newton-cg. The Grid SearchCV on Logistic Regression classifier increased the ROC-AUC score to 81%.

Results

The project proposal shows that employee attrition can be predicted using machine learning model. Different models like Logistic Regression, Decision Tree Classifier, SGD classifier, Random Forest was evaluated. Logistic Regression model has best score. The model was evaluated using test data and score of tuned model is 78.90%.

Figure 10 shows the confusion matrix where bottom right is True Positive, top left is True Negative, top right is False Positive and bottom left is False Negative.

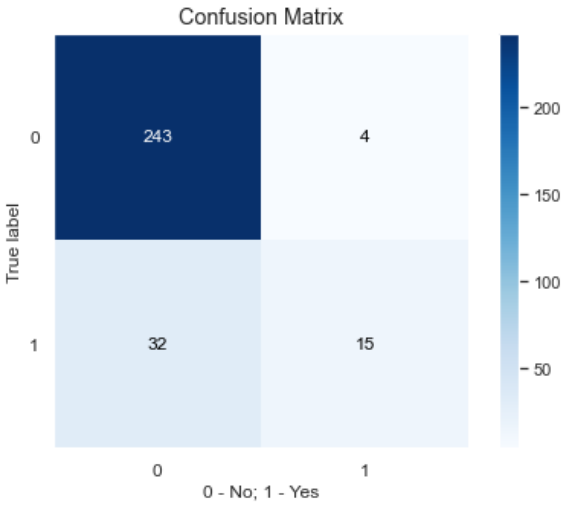

Figure 10

True positive rate is 79% which indicates attrition was predicted correctly when the data indicated attrition. True negative rate is 88% which indicates model correctly predicted employee would stay with company when data indicated non-attrition. False positive rate is 21% which indicates the percentage of employee that were predicted as who would leave the company but was actually employee who stayed. False negative rate is 11.63% which indicates the percentage of employee that were predicted as who would stay but they actually left.

Organizations can use this tool to determine employee attrition and focus on the reasons that affect attrition. With this, employee can stay longer with the organization which reduces additional cost to the company and reduce potential delay in organization commitments.

References

1. Alexandra. (2021, July 6). Employee attrition vs. employee turnover: What are these metrics about? Harver. https://harver.com/blog/employee-attrition-employee-turnover/.

2. toolbox.com. (n.d.). https://www.toolbox.com/hr/engagement-retention/articles/what-is-attrition-complete-guide/.

3. 8 essential employee Retention Factors modern EMPLOYERS IGNORE. Rise. (2021, June 24). https://risepeople.com/blog/employee-retention-factors/.

4. 5 factors that lead to high employee turnover. Fingercheck. (2020, August 24). https://fingercheck.com/5-factors-that-lead-to-high-employee-turnover/.

5. Why employees stay. Harvard Business Review. (2014, August 1). https://hbr.org/1973/07/why-employees-stay.

6. Zojceska, A. (2020, April 3). HR metrics: How and why to Calculate employee turnover rate? Blog. https://www.talentlyft.com/en/blog/article/242/hr-metrics-how-and-why-to-calculate-employee-turnover-rate.

7. Heinz, K. (n.d.). 40 employee turnover statistics to know. Built In. https://builtin.com/employee-turnover-statistics.

8. Attrition rate: What you need to know. Personio. (2021, July 29). https://www.personio.com/hr-lexicon/attrition-rate/.

9. Kuepers, J. (2021, July 1). Employee turnover: Why fixing it now is urgent. Click Boarding. https://www.clickboarding.com/employee-turnover-what-is-it/.

10. Employee turnover rates: An industry comparison. Edays. https://www.e-days.com/news/employee-turnover-rates-an-industry-comparison.

11. King, S. (n.d.). More than you think: The cost of employee turnover. Outsourced Bookkeeping, Accounting and Controller Services for Small Businesses and Nonprofits. https://www.growthforce.com/blog/the-real-cost-of-employee-turnover-its-more-than-you-think.

Questions

What are the common factors that affect employee attrition?

Some factors are age, overtime, total working years, years with current manager.

What measure can organization take to reduce attrition?

Organizations can take measures to look at specific age group, look at their prior

experience and address it accordingly.

Is attrition more in younger age group?

Yes, Attrition is more in age group 25 to 35.

Does employee salary affect attrition?

Yes, attrition is more for monthly income between $1500 and $3500

What is the error rate of the model proposed?

False Positive rate is 21% and False Negative rate is 12%.

Is there any feature that does not impact attrition?

Performance Rating and Business Travel are few features with less correlation

Does employee role play any part in attrition?

Attrition is more among job role – Laboratory Technician, Sales Executive, Research Scientist and Sales Representative

Do employee leave organization when they work overtime?

Yes, employee tend to leave when they are working overtime.

What is the top reason for attrition?

Top reasons are overtime and total working years

Can the model prediction be further improved?

Yes, it can be improved further when analyzing with more data values.