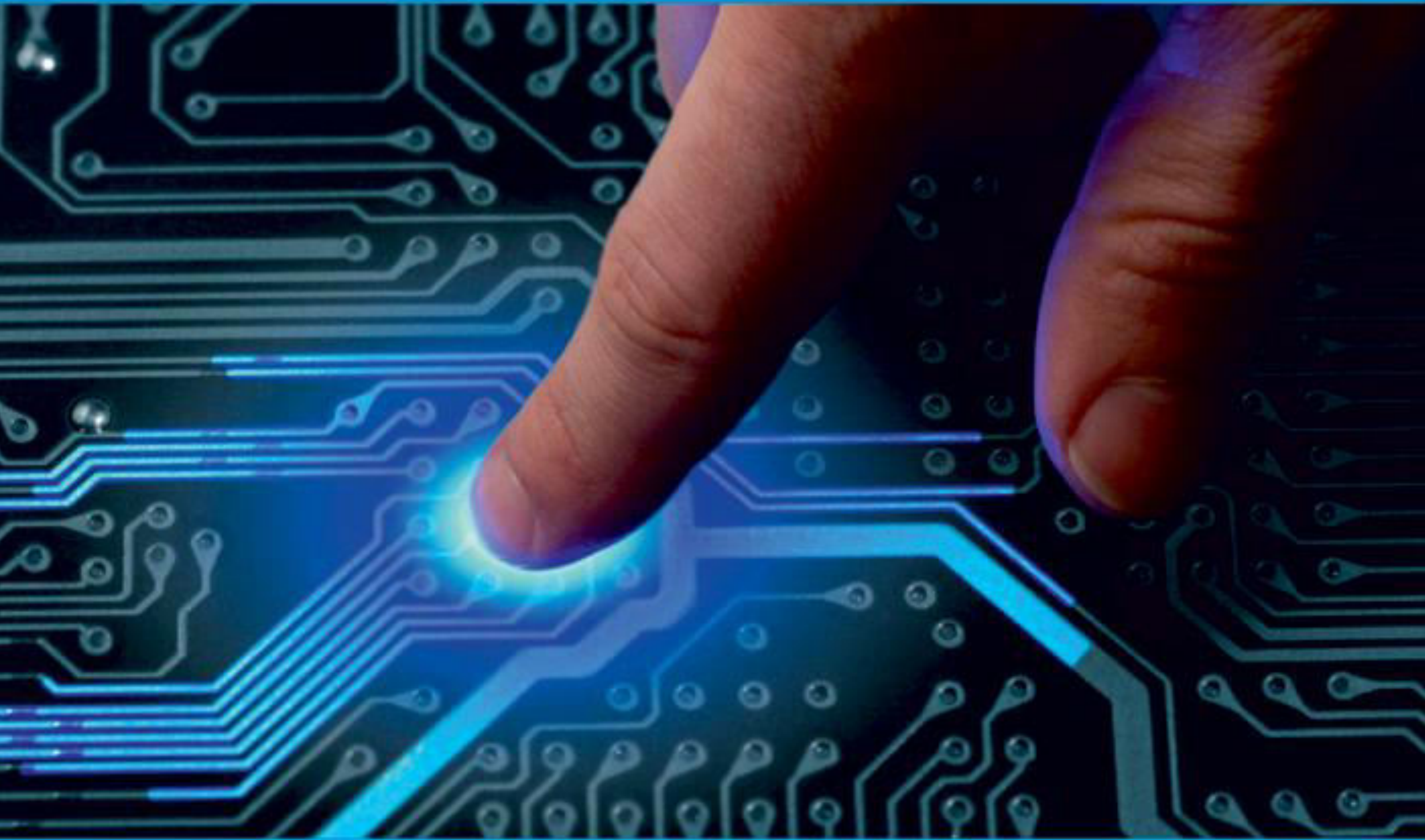
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| DOI: 10.15680/IJIRCCE.2022.1004007 |

Employee Attrition & Turnover Prediction using Machine Learning

Henil Jain, Pujan Kothari, Jainam Shah, Vashali Chavan

Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India

ABSTRACT: Employees are the most valuable assets an organization has. Organizations invest a lot of time and resources in recruiting and training employees, depending on their strategic needs. When an employee leaves a company, the organization loses not only the key employee but also the resources invested in hiring and selecting those employees and training them in their related tasks. So, In this paper, we will develop machine learning models to predict the attrition rate using the HR analytics dataset. Also to anticipate the reasons for valuable employees to leave and the undiscovered territories where companies are lagging. These models benefit HR management to mend their methodologies to make certain decisions and to consider important factors before recruiting.

KEYWORDS: Attrition, supervised learning, classification, regression, feature selection, data cleaning, Logistic regression, random forest, Adaboost, Xgboost, Human resource, Naives bayes, SVM

I. INTRODUCTION

A decrease in manpower in any organization due to workers leaving volunteering or due to resignation, and not being replaced is known as Employee Attrition. Employee Turnover is the ratio of the number of existing employees traded by a set of new employees over time. Companies with a high Attrition rate or vacancy of employees are proportional to high employee turnover. This vacancy of workers disrupts the functioning of the organization. This leads to an item of huge expenditure on human resources, for new enrolment and to train freshly hired recruits likewise in the performance management. Therefore, in any organization Attrition is unavoidable and it causes inadequate performance. By estimating the work environment, improving employees' morale and under desirable working conditions, we can explicitly diminish this issue. Employees leaving companies have different explanations. So, by predicting the attrition level and its factors, it will enable the management to act better by upgrading their methodology and strategies. When a valuable and talented employee is on the urge of leaving, offering a few recommendations like a better training program or a pay increment, incentives reduce the chances of attrition. In the modern era, Machine learning models to anticipate attrition value are highly demanding and thriving. These models make decisions for Human Resource Management with strategies by considering all the factors. A connection between features of both active and terminated employees is developed thus, formulating companies with a better structure which improves their efficiency, functioning and reduces the employee turnover.

II. REVIEW OF LITERATURE

[1] Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca predicted Employee Attrition Using Machine Learning Techniques in 2020. They statistically assess the data and classify them for selective classification algorithms to evaluate the algorithm's performance. The highlight of their research was to emphasize feature selection and their relations like delay and absenteeism, distance on work turnover in thriving segments. [2] Jayashree Prasad, Prathamesh Shinde, Manodnya Gaikwad, Tanmay Mahindrakar, Sohan Kakatkar exposition on Employee Attrition Prediction in 2020. They utilized prescient data procedures like the decision tree approach and a logistic regression model for analyzing turnover by illustrating it on a real-life data set. The use of the Sequential Backward Selection Algorithm (SBS) removed less significant attributes in feature engineering and second level Chi-square which identified the significance of each of the attributes. Following, An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection by [3] Saeed Najafi-Zangeneh, Naser Shams-Gharneh, Ali Arjomandi-Nezhad, and Sarfaraz Hashemkhani Zolfani issued in 2021. They Explicitly developed a “ max-out “ feature that reduces the dimension of feature space to improve accuracy. Also, the application of the logistic regression method to build up a risk equation. Later this equation was applied to assess attrition risk with the current set of employees. After the estimation, a high-risk cluster was recognized to discover the reasons and henceforth an action plan was selected to minimize the risk.

[4] Fredrik Norman works on Predicting employee attrition with machine learning on an individual level, and the effects it could have on an organization in 2020. utilization of three prescient data mining procedures (logistic regression, classification and regression trees, classification trees) on a sample data of 150 employees in a large software organization. They used a random forest algorithm and had an accuracy of 88%. It demonstrates a connection between withdrawal behaviors and employee turnover. Over a larger scale, variables can be collected.

III. Algorithms

In this paper, we talk about both types of algorithms which are supervised algorithms to find and chart the best model with greater accuracy for our working

Logistic Regression:

Logistic Regression is a classification model which fits the data values into a logistic function. It is used to convert the values into a categorical or a discrete form i.e ranging between 0 to 1.
The general equation of logistic regression is represented by-

$$P\ (\ Y\,|\,X,W)=1/(1+e\text{\textasciicircum}(\ e\text{\textasciicircum}(--(w0+\Sigma\ wixi))$$

In Logistic regression, instead of fitting a regular linear regression line, it fits the data in an ‘S-shaped logistic function, which predicts one maximum and one minimum value. A logistic regression model is used for solving classification problems.

Decision Tree:

Decision tree is a supervised learning algorithm where the input data is split continuously according to the parameters mentioned. It has a tree-like structure classifier, where the internal nodes denote the features of a given dataset. Each leaf node denotes the outcome and branches denote the decision rules. It is named a decision tree because of its similarities like a tree, where it starts with a root node and the root nodes further expand into branches. The decision tree begins with an original set and then calculates the Entropy(H) and Information Gain(IG) of each Iteration.

The General Equation of Entropy(H) is given as -

$$E=-\sum pilog2pi$$

Random Forest:

Random Forest is a supervised learning algorithm where it creates different decisions to portray different samples and takes an average for regression and bulk vote for classification. It can be used to solve both regression and classification problems. It is useful for handling data sets that have continuous values for regression and categorical or discrete values for classification. It is formed on the concept of ensemble learning where it combines multiple classifiers to solve complex problems. More the numbers of trees in the forest will lead to higher accuracy of the model. Random Forest can maintain its accuracy when a large part of data is also missing leading to consistency.

Adaboost Classifier:

The AdaBoost Classifier, which is also short for Adaptive Boosting, is a Boosting method that is used as an Ensemble Method. It is called Adaptive Boosting as the weights are redistributed in each case, with higher weights given to incorrectly planned conditions. Boosting is used to reduce bias and variability in supervised learning. It works on the principle that learners grow in succession. Except for the first case, each subsequent learner is extended from previously grown learners. In simple terms, weak learners are transformed into strong ones. The AdaBoost algorithm works on the same principle as upgrading with small differences.

The mathematical formula for the Adaboost classifier is shown below,

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Xgboost:
XGBoost is an ensemble type of supervised learning algorithm which means that it mixes the results of many simple, weak models, called base learners to make a single result. It makes use of gradient boosted decision trees. It can be used to solve regression, classification, and user-defined prediction problems.

Naive Bayes:
The Naive Bayes algorithm is a supervised learning algorithm that is based on the Bayes theorem. It uses the assumption that all variables are independent of each other, and after that, it calculates all possibilities that are used to differentiate. The algorithm works as follows: to get the Y output function given a set of input variables X, the algorithm estimates the values of P (X | Y) and P (Y), and then uses Bayes' rule to calculate P (Y | X), which is the required result, for new samples.

Support Vector Machine(SVM):
Support Vector Machine or SVM is a supervised Learning Algorithm. It can be used to solve both types of problems i.e classification and regression. In this algorithm, we arrange each data object as a point in space n (their n number of features you have) per value of an element that is the value of a particular combination. Then, we make a distinction by finding a hyperplane that separates the two categories.

IV. RESEARCH METHODOLOGY

We use a broad range of machine learning techniques ranging from as simple as Logistic regression, Naive Bayes to more complex techniques such as SVM or Support Vector Machine, Xgboost and Random Forest.

A. Basic Working/Architecture

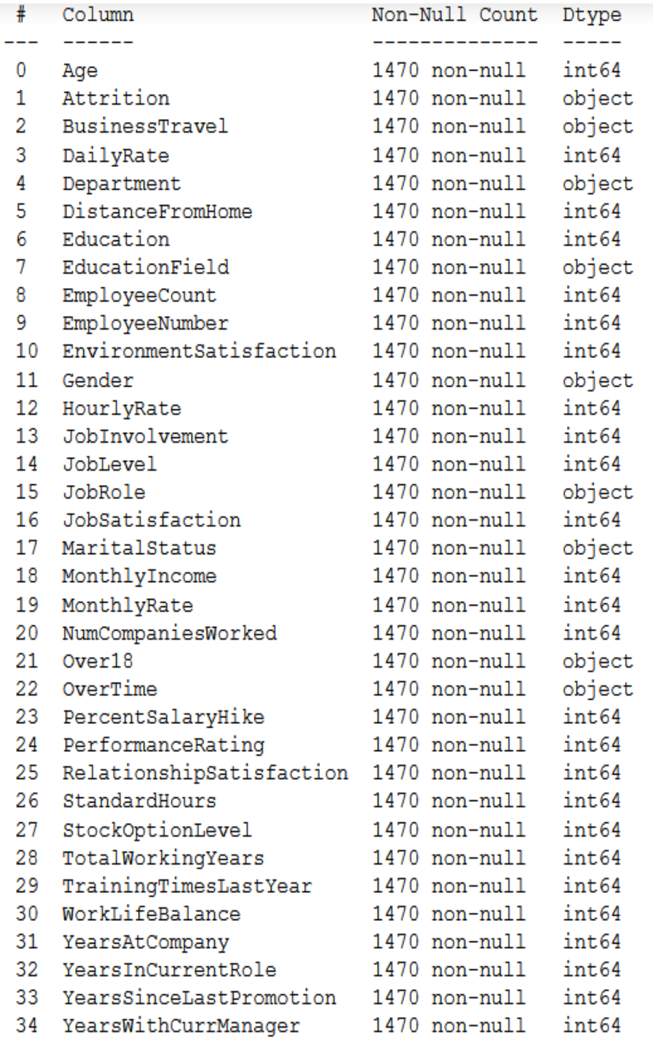
- Identify an employee data set that includes current and past employee records
- Clean the database, manage missing data and determine new features whenever needed.
- Remove outliers and Null values from Dataset and apply a logistic function to minimize the range.
- Select features among employee data that are appropriate for the attrition prediction.
- Use the feature selection method using bins, then select more convenient features to predict employee attrition in companies.
- Try applying classification, regression algorithms and store the precision, recall, Accuracy and F-measure results.
- Compare the results of baseline models and find out the best model to predict employee attrition by comparing each other with their precision, recall, accuracy and f-measure results.
- Apply hyperparameter tuning on the selected model to increase accuracy and precision even more.

B. Data Set Description

In this study, we have used publicly available dataset, which can be found on IBM Watson Analytics1. The database includes the engineer information generated by IBM data scientists. The database contains HR-related information for 1470 employees with 34 attributes. The attributes had details about employee age, Distance from home, Education field and many more Also, a total of 1233 active workers were coming out of the “No” attrition category though the remaining 237 employees were from "Yes" Categorized. In this study, Three points were removed: 'Number of employees', because the sequence of numbers (1,2, 3.); 'normal hours', for which all employees have the same regular hours and Lessthan18 as everyone working for a company is above 18.. Also, all non-numbers numbers are assigned to numerical values for such processing such as: Sales = 1, Research & Development = 2, Human Resources = 3.

| DOI: 10.15680/IJIRCCE.2022.1004007 |

Tab1e I. HR data set attributes



```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
```

C. Proposed Model

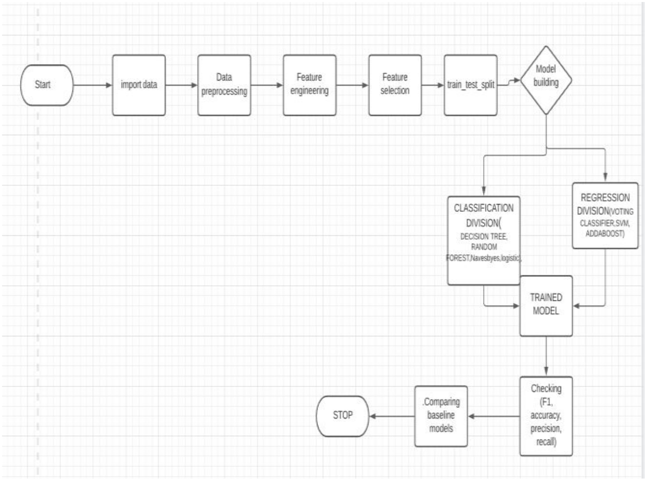

Start
import data
Data preprocessing
Feature engineering
Feature selection
train_test_split
Model building
CLASSIFICATION DIVISION(DECISION TREE, RANDOM FOREST,Navesbyes,logistic)
REGRESSION DIVISION(VOTING CLASSIFIER,SVM, ADDABOOST)
TRAINED MODEL
Checking (F1, accuracy, precision, recall)
.Comparing baseline models
STOP

Fig 1: Proposed Model

IV. DATA ANALYSIS

A. Data Pre-processing

From the IBM employee database, we apply a feature selection method where we choose the most important features of the dataset and then divide the total database into two smaller datasets. One dataset is the test dataset and another dataset is the training dataset.. Before dividing the datasets, we make sure any feature in the record does not contain any null value or value that is not defined or insignificant. We also check for outliers and remove any outliers using a log function to remove any redundancy. Then we create bins for our features to get them in categorical values such that it is easier for our models to understand their significance. After we create a perfect dataset, we divide them into training and test datasets. The test database contains all the important features for employee attrition prediction.

B. Test and Train Dataset

Dividing data into test data sets and training data sets is an important part of testing data mining models. With this separation of the amount of data set into two data sets we can reduce the effects of data inconsistencies and better understand the features of the model. The test data set contains all the necessary data prediction data and the training data set contains all non-essential data. Here we have 788 records in the test database and 682 records in the training Database. We use data classification and predict the test databases of 788 records.

C. Data Visualization

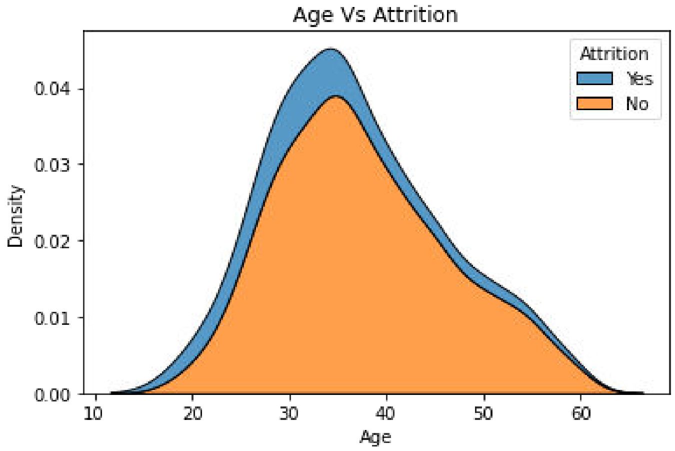


Fig 2: Age Vs Attrition

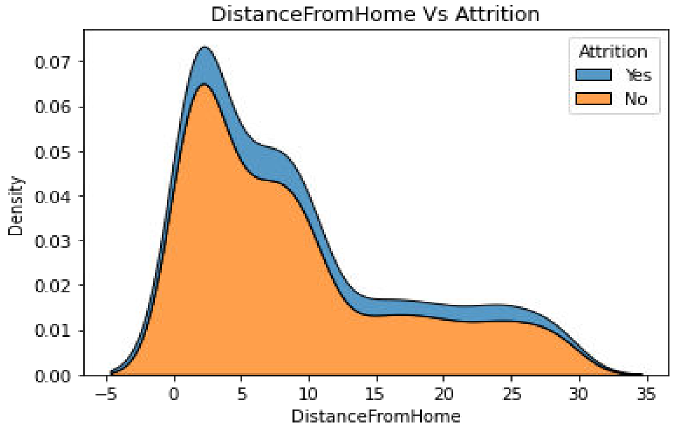


Fig 3: Distance From Home Vs Attrition

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004007 |

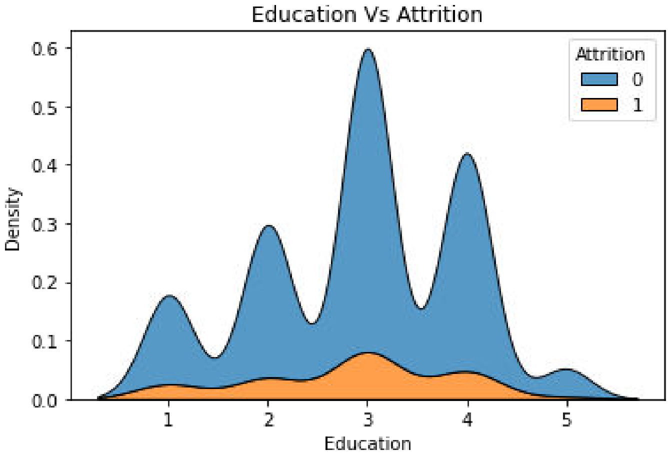


Fig 4: Education Vs Attrition

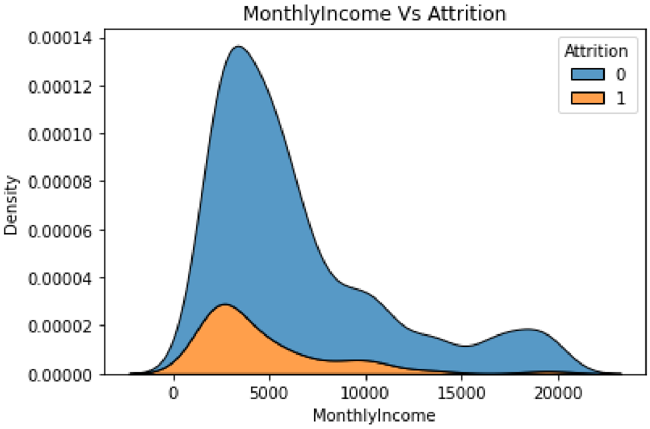


Fig 6: MonthlyIncome Vs Attrition

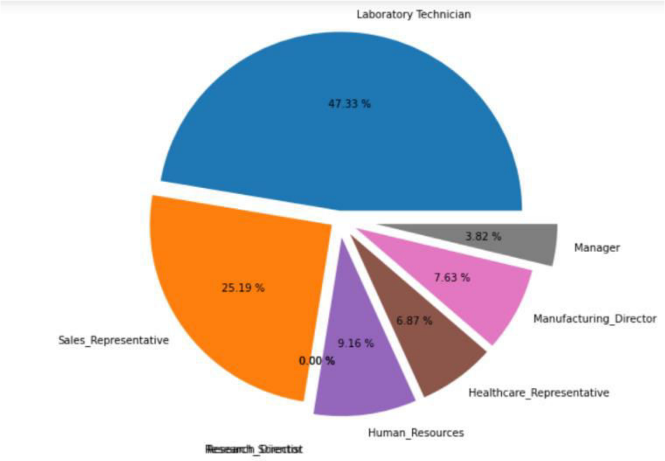


Fig 8: Job Wise Attrition

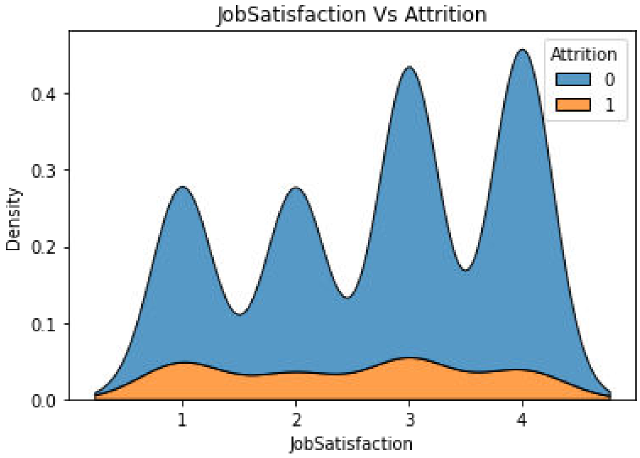
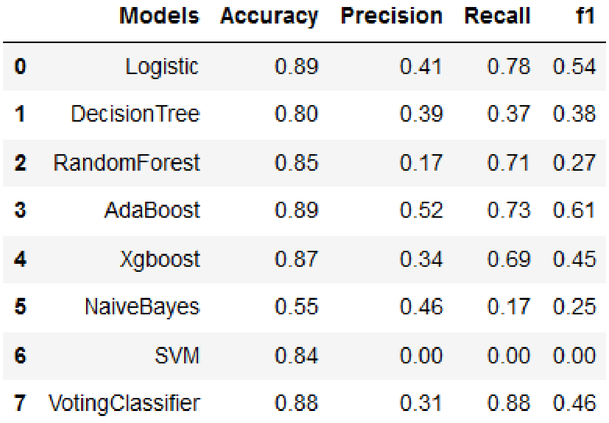


Fig 7: Job Satisfaction Vs Attrition

VI. Result

Table II. Results of different models



	Models	Accuracy	Precision	Recall	f1
0	Logistic	0.89	0.41	0.78	0.54
1	DecisionTree	0.80	0.39	0.37	0.38
2	RandomForest	0.85	0.17	0.71	0.27
3	AdaBoost	0.89	0.52	0.73	0.61
4	Xgboost	0.87	0.34	0.69	0.45
5	NaiveBayes	0.55	0.46	0.17	0.25
6	SVM	0.84	0.00	0.00	0.00
7	VotingClassifier	0.88	0.31	0.88	0.46

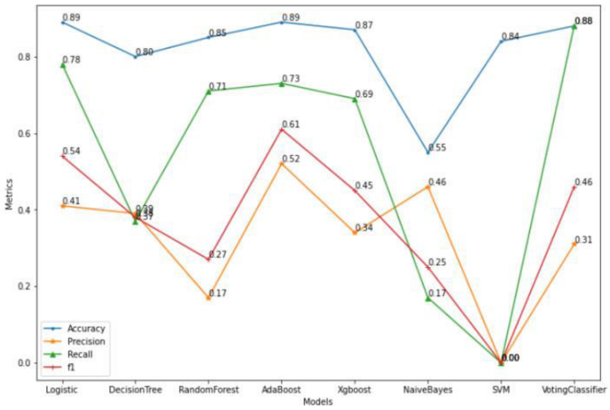


Fig 9. Comparing Baseline models

VI. CONCLUSION

In this research we have built very simple models for predicting employee attritions, from some basic Test Data Analysis to feature engineering and the use of the Logistic regression and Adaboost classifier which both give an Accuracy of 89%. Furthermore, Adaboost Classifier has a higher precision of 52% out of all models used to predict Employee attrition. The most common explanation for employee attrition is possibly the effort-reward inequality. In this case, this usually applies to the people who are at work longer than necessary and employees who have for the most of the time had a very low pay. It should be checked whether there is a compelling strategy for overtime in the organization. We have also found that various aspects of the working life balance may address the issue of our representatives. How the three elements are connected (directly or indirectly) to working life balance (distance from home, business travel, and work life all things to consider) have their place among the top 20 features; it may be an indication that something must be done around there. If we take our "Research" as an example of IBM current employees, we can say that the role of Laboratory technician has a very high probability of attrition rate. We can definitely do something about it and try to reduce the attrition rate.

REFERENCES

[1] Khaled Alshehhi, Safeya Bin Zawbaa, Abdullah A Abonamah, Muhammad Usman Tariq, " EMPLOYEE RETENTION PREDICTION IN CORPORATE ORGANIZATIONS USING MACHINE LEARNING METHODS ", Volume 27, Special Issue 2, 2021, IEEE paper

[2] Jayashree Prasad, Prathamesh Shinde, Manodnya Gaikwad, Tanmay Mahindrakar, Sohan Kakatkar, "Employee Attrition Prediction," 2020 International Journal of Future Generation Communication and Networking, Vol. 13, No. 3s, (2020), pp. 724–730

[3] Francesca Fallucchi, "Predicting Employee Attrition Using Machine Learning Techniques", Review of Business Research, 2020, Research Gate.

[4] Krishna Sehgal, Harlieen Bindra, Anish Batra, and Rachna Jain, "Prediction of Employee Attrition Using GWO and PSO Optimised Models of C5.0 Used with Association Rules and Analysis of Optimisers", Springer Singapore, vol.74, June 2019.

[5] S. K. Monisaa Tharani and S. N. Vivek Raj, "Predicting employee turnover intention in IT&ITeS industry using machine learning algorithms," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), 2020, pp. 508-513, DOI: 10.1109/I-SMAC49090.2020.9243552.

[6] S. N. Khera and Divya, "Predictive Modeling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques," Vision, vol.23, no. 1, pp. 12–21, Mar. 2019, DOI: 10.1177/0972262918821221.

[7] Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, Xiaoyu Zhu, "Employee turnover prediction with machine learning: a reliable approach", Springer Nature Switzerland AG, vol. 869, 2019.

[8] Sri Ranjitha Ponnuru Gopi Krishna Merugumala, Srinivasulu Padigala, Ramya Vanga, Bhaskar Kantapalli, " Employee Attrition Prediction using Logistic Regression," MAY 2020 International Journal for Research in Applied Science & Engineering Technology (IJRASE)2321-9653 IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue V

[9] Saeed Najafi-Zangeneh, Naser Shams-Gharneh, Ali Arjomandi-Nezhad, and Sarfaraz Hashemkhani Zolfani, "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection," 2021 MATHEMATICS MDPI JOURNAL, 2014, pp. 273-278, DOI: 10.1109/SOCPAR.2014.7008018...

[10] Sri Ranjitha Ponnuru Gopi Krishna Merugumala, Srinivasulu Padigala, Ramya Vanga, Bhaskar Kantapalli, " Employee Attrition Prediction using Logistic Regression," MAY 2020 International Journal for Research in Applied Science & Engineering Technology (IJRASE)2321-9653 IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue V

[11] Abhishek Sethy1, Dr. Ajit Kumar Raut2, " EMPLOYEE ATTRITION RATE PREDICTION USING MACHINE LEARNING APPROACH ", December 2019 ISSN 2651-4451 | e-ISSN 2651-446X

[12] Anjali Chourey, Prof. Sunil Phule, Dr. Sadhna Mishra, " A SURVEY PAPER ON EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING TECHNIQUES ", Journal of Interdisciplinary Cycle Research, Volume XI, Issue XII, December/2019

[13] Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Sayali Sutar, Rajarshi Sadafule, Suhasini Bhat, " Employee Attrition Predictive Model Using Machine Learning ", Volume: 07 Issue: 05 | May 2020 e-ISSN: 2395-0056

[14] Rachna Jain, Anand Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach" Publisher: IEEE, Date of Conference 24th November 2018

[15] Sheik N. Khera, Divya "Predictive Modeling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques" Published March 2019

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING