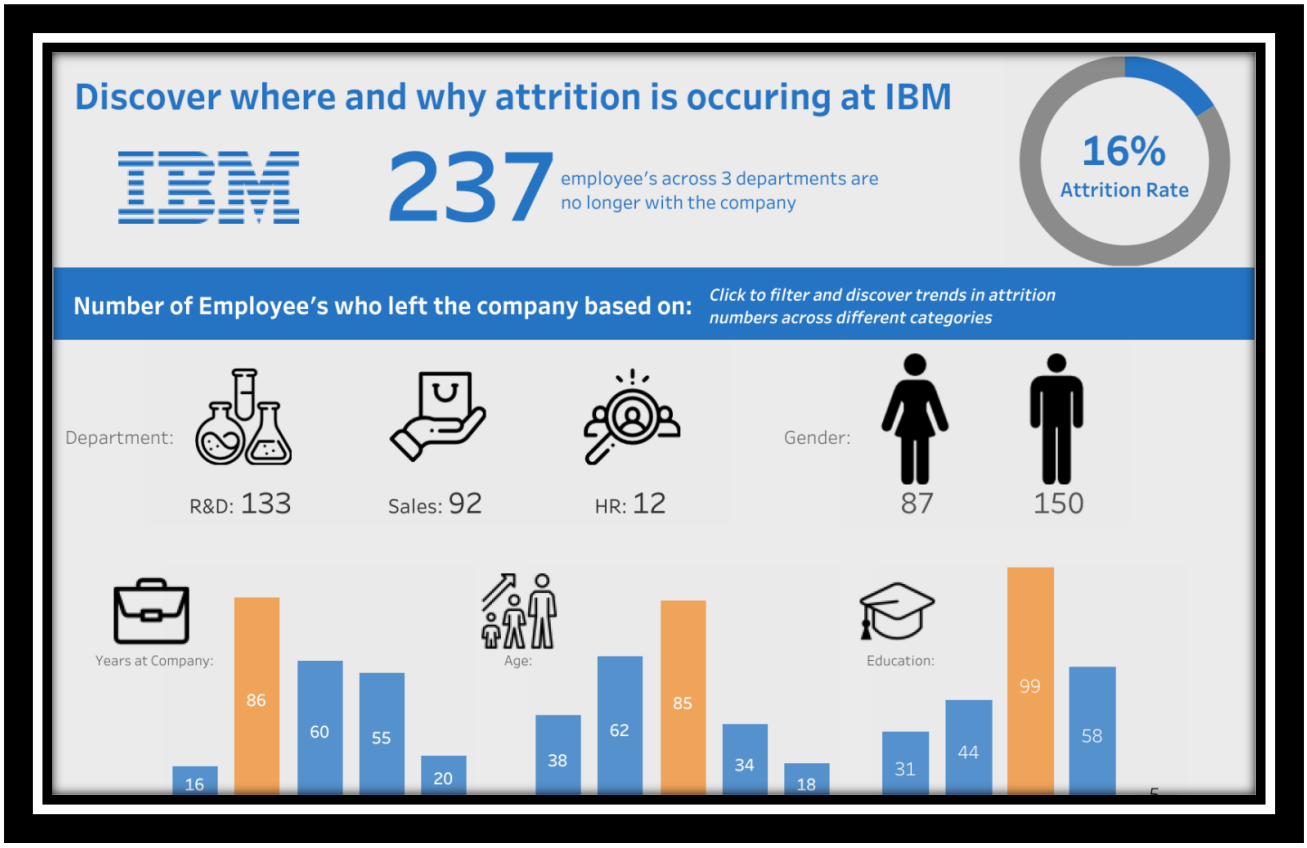
HR ANALYTICS PROJECT



Submitted by:

UJJWAL PRATIK

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT ATTRITION

Employee attrition is when an employee leaves the company through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or even illness or death. Whenever anyone ceases working for the company for any reason and is not replaced for a long time (if ever), that would be employee attrition.

PROBLEM STATEMENT

Attrition in human resources refers to the gradual loss of employee's overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees. How does Attrition affect companies? and how does HR Analytics help in analysing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by steps.

ATTRITION AFFECTING COMPANIES

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

'Age', 'Attrition', 'Business Travel', 'Daily Rate', 'Department',
'Distance From Home', 'Education', 'Education Field', 'Employee Count',
'Employee Number', 'Environment Satisfaction', 'Gender', 'Hourly Rate',
'Job Involvement', 'Job Level', 'Job Role', 'Job Satisfaction',
'Marital Status', 'Monthly Income', 'Monthly Rate', 'Number Companies Worked',
'Over18', 'Over Time', 'Percent Salary Hike', 'Performance Rating',
'Relationship Satisfaction', 'Standard Hours', 'Stock Option Level',
'Total Working Years', 'Training Times Last Year', 'Work Life Balance',
'Years At Company', 'Years in Current Role', 'Years Since Last Promotion',
'Years With Current Manager'

sample

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7

OBJECTIVE

- Our main objective is to predict and Understanding the Attrition in HR using machine learning algorithms.
- All the parameters will be analysed through Machine Learning algorithms like Logistic Regression, AdaBoost Classifier, Random Forest Classifier, Decision Tree Classifier, Gaussian NB, Support Vector Classifier etc which will help to predict the Attrition.

DATA DESCRIPTION

- The source of data is taken from GitHub.
- Data Link is (https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics).

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the attrition because it gives the best outcome.
- The dependent variable is Attrition.

METRIC USAGE

a. Logistic Regression.
b. AdaBoost Classifier.
c. Random Forest Classifier.
d. Gaussian NB
e. Decision Tree Classifier.
f. Support Vector Classifier

SYSTEM REQUIREMENTS

Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- Logistic Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- AdaBoost Classifier: s a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
- Random Forest Classifier: The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.
- Gaussian NB: Gaussian NB is based on the Naive Bayes theorem with the assumption of conditional independence between every pair of features given the label of the target class.
- Decision Tree Classifier: A decision tree is a flowchart-like tree structure in which the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. A Decision Tree consists of Nodes: Test for the value of a certain attribute.
- Support Vector Classifier: Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. In addition to this, an SVM can also perform non-linear classification.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Zscore: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- IMB Learn Anaconda: imbalanced-learn is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. It is compatible with scikit-learn and is part of scikit-learn-contra projects.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Label Encoding
c. IMB Learn
d. Logistic Regression
e. AdaBoost Classifier
f. Gaussian NB
g. Decision Tree classifier
h. Support Vector Classifier
i. Grid Search CV
j. Cross Validation

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Compared between two columns.
- Checked and removed outliers through zcore method.
- Removed skewness present in the dataset.
- Done Oversampling.
- Cross validate the accuracy score from overfitting.
- Hyper Parameter tuning.

CONCLUSION

As our conclusion we proclaim that, after checking accuracy score, cross validation, Ensemble Techniques, and checking AUC score we declare AdaBoost Classifier predicting 84.87% accuracy is best suited model for our purpose of predicting Attrition in HR.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.