Predicting Employee Attrition

Problem Statement:

In recent years, attention has increasingly been paid to human resources (HR), since worker quality and skills represent a growth factor and a real competitive advantage for companies. After proving its mettle in sales and marketing, artificial intelligence is also becoming central to employee-related decisions within HR management. Organizational growth largely depends on staff retention. Losing employees frequently impacts the morale of the organization and hiring new employees is more expensive than retaining existing ones.

You are working as a data scientist with HR Department of a large insurance company focused on sales team attrition. Insurance sales teams help insurance companies generate new business by contacting potential customers and selling one or more types of insurance. The department generally sees high attrition and thus staffing becomes a crucial aspect.

To aid staffing, you are provided with the monthly information for a segment of employees for 2016 and 2017 and tasked to predict whether a current employee will be leaving the organization in the upcoming two quarters (01 Jan 2018 - 01 July 2018) or not, given:
1. Demographics of the employee (city, age, gender etc.)
2. Tenure information (joining date, Last Date)
3. Historical data regarding the performance of the employee (Quarterly rating, Monthly business acquired, designation, salary)

Data Dictionary

Variable	Definition
MMMM-YY	Reporting Date (Monthly)
Emp_ID	Unique id for employees
Age	Age of the employee
Gender	Gender of the employee
City	City Code of the employee
Education_Level	Education level : Bachelor, Master or College
Salary	Salary of the employee
Dateofjoining	Joining date for the employee
LastWorkingDate	Last date of working for the employee
Joining Designation	Designation of the employee at the time of joining
Designation	Designation of the employee at the time of reporting
Total_Business_Value	The total business value acquired by the employee in a month (negative business indicates cancellation/refund of sold insurance policies)
Quarterly Rating	Quarterly rating of the employee: 1,2,3,4 (higher is better)

Solution and Approach

Understanding Of Data

Employees leaving the organization depend on lots of factors. There can be internal and external factors. We can find those factors from the data using Machine Learning Algorithms.

Data says a lot but it depends on how a person approaches that data.

Let's see some factors which lead employees to leave the organization. There can be countless reasons but these are some frequent one.

1. Needing more of a challenge
2. Looking for a higher salary
3. Feeling uninspired
4. Wanting to feel valued
5. Seeking a better management relationship
6. Searching for job growth and career advancement
7. Needing more feedback or structure
8. Wanting a different work environment
9. Looking to live somewhere else
10. Feeling conflicted with workplace policies
11. Thinking that their job has changed
12. Wanting a clearer company vision
13. Needing a better work-life balance
14. Seeking a more financially secure company
15. Wanting more independence
16. Looking for more recognition

Let's try to find these insights from the given data.

Understanding of data is one of the critical steps before starting any Problem. If We understand the Data then we can apply all necessary transformations.Write Preprocessing of data can solve your problem to 60-70% then applying algorithms is not a big task.

Steps to solve the Problem:

Step 1: Finding relevant columns i.e. Feature Reduction from the data.

These are some important columns according to me.

Columns_Name = ['Age', 'Gender', 'City', 'Education Level', 'Salary', 'Date of joining', 'Joining Designation', 'Designation', 'Total Business Value', 'Quarterly Rating']

"**Age**" shows how many employees have been in the industry and these can be a factor where employees will start looking for better opportunities.

"**Gender**" also impacts employees leaving the job in India. It will not affect other countries. I consider it since we live in India.

"**Education Level**" also impacts since if the level is low then people leave company to pursue Higher Education.

"**Salary**" is one of the main factors for employee attrition since people leave the company to get better paid.

"Date of Joining" shows how many years a person in the company and employee may be looking for a fresh start.

"**Designation**" shows if employees get recognition or salary hike. Since this matters a lot.

"**Total Business value"** shows how effective the employee is for the organisation and if it is lower than he might think not getting the right work.

"Quarterly Rating" shows the performance of employees. If it is low then he will try to change so that performance might improve.

Step 2: Creating the New Training Dataset.

Since Train data contains unstructured time series data where we have unstructured monthly Employee ID data. Employee ID should be unique so that we can build an efficient model.

Grouping the data, rightly can do the task. I group the data based on Emp_ID and take the average of all the other columns. I also ensure that data should not contain any bias.

I also add a new target column as employee_leave_or_not using column Last Working data.

If Last Working data has blank then it is 0 if it has value than 1.

Step 3: Pre-processing of Data.

Since data was imbalanced I used a rescale method to remove the imbalance from data.

Convert categorical data i.e. City, Gender, Education level, Joining date into numerical, since ML models don't understand the categorical data.

Standardise or scale the data so that all the values lie between 0 and 1.

Remove all the unnecessary data and do the data type conversion.

Step 4: Split the data and train the model

Splitting the data into train and test to check the accuracy of Algorithm.

Get Train variables and target value.

Apply all the Algorithms which are available to compare various algorithms accuracy.

Step 5: Hyper tuning the model

After getting accuracy check the confusion matrix and F-1 score.

Do hyper tuning of the model by changing the parameters passed to the model to get better accuracy.

Step 6: Uploading of files

Zip the code and approach file and generate the .csv file.

Upload the files to the portal.