BITGRIT ROUND 2 REPORT

REPORT OF ANALYSIS OF DATASET

“Team -D” IIT (BHU) Varanasi

Introduction

Dataset given was of a sales organisation, which contained information related to attrition from the company with the information in the form of ID, grade, age, tenure etc. of the employees undergoing attrition. The analysis was done on the dataset, various graphs, boxplots etc. were made to gain the insights, relation between features, the behaviour of each feature individually and according to other features. Some insights were drawn and the importance of features was identified. Some features were converted to a readable format by the model. Duplicacy (same value with different name [lowercase/uppercase error] occurring differently) was removed from the features. Thus cleaning of data, analysis, drawing insights from it and finally building a predictive model was done.

Possible Use Cases

From the data given, the following use cases were visible.

1. **Attrition prediction with the information of the employee (yes/no)[whether he/she will leave the job or not]**
2. **Forecast of the number of attritions according to month(based on In Active date)].**
3. **Tenure prediction using the following features:**

   Salary, Grade, Education, Zone, Marital Status, Gender and Age when joined (Date of joining-tenure[for training])

Work was done mainly on the first use case but a glance was put on the 3rd use case also.

Detection of categorical and continuous features

The categorical and continuous features were identified with their value counts. After doing the value counts, the following features were found to be categorical:

:- Designation, Grade, Location, Gender, Education, Last Rating, Marital Status, Zone, Remarks

And these were found to be continuous:

:- DOJ, Age, Monthly income, Engagement score.

Cleaning of the Data

- The Designation and Zone features were two categorical features, having an impurity that same value occurred in different ways example south occurred as SOUTH and south and pandas considered it as two different values. This duplicacy was handled.
- Tenure being a continuous column was read as object data type, 1.11 which means 1 year and 11 months was read as 1.11 year thus 1.11 was considered less than 1.9, for removing this, a new feature was added known as ‘total_months’ which displayed tenure in no. of months worked.
- %satisfaction was converted from string to float.

Some insights from the data

1. Designation and Grade plots are exactly the same, explaining the grades of the designation.
2. Sales executives leave the job most, followed by sr. sales executive :   for this reason can be as follows:

a. Sales Executives and Sr. Sales executives are at employee grade, they are not at a managerial post.
b. Both of these grades are maybe in search of an opportunity providing some managerial work.
c. As we go towards more managerial part of the company, attrition amount decreases.

3. Gender wise attrition is widely unbalanced, 5.5% Females and 94.5% of males leave the job:
Some reasons for this can be as follows:
a. Maybe there is a very unbalanced gender ratio among the people working in the company.
b. Or Females are more prone to work in an organisation without changing it, i.e. they crave for the stability of the job.
4. Education-wise attrition is widely unbalanced too, 94.5% B.Tech, 5.5% MBA Graduates
Some reasons for it can be as follows:
a. MBA Graduates usually go to the managerial jobs, i.e. the jobs with higher managerial grades, and as their attrition is low, thus MBA grads are less prone to attrition.
b. B tech graduates are freshers, they seek for growth and opportunity, which they get when they make a switch, thus, they opt for more attrition.
5. Last Rating wise attrition shows that the people who were rated good according to their work, are more prone to attrition.
Reasons for it can be as follows:
a. Those who are rated good, are more prone to make a switch as they consider themselves suitable for a better position.
b. But the ones who are exceptionally rated 5 i.e. best are less prone to attrition, reason can be that they have started liking their job, thus are great at work and having good ratings.
6. Marital Status does not give a good insight of the matter, we have to dig deep into it for better insights
7. Zone wise distribution, gives a good insight of attrition, stating that South Zone constitutes to 30.6% of attrition, followed by North zone, constituting 22% of total attrition. Thus for finding the reason behind we have to dig deep into it.
8. Remarks column indicated that, 'issues with manager' is the reason given by most for attrition, followed by 'lack of growth' which was intuitive as many who left the job, were rated 4/5 thus being capable they would search for better position which might not have been possible in the past situation when they were employees.
9. "ho" zone contains more attrition of higher-order jobs.
10. "South" zone contains more attrition of lower order jobs.
11. Termination due to poor performance is mainly in lower graded jobs like e1 and e2.
12. Lack of growth as a reason for attrition decreases as we go towards higher graded jobs.
13. Higher-order jobs' employees seek more challenging job roles/higher designation thus they leave the job.
14. From "ho" zone, an MBA graduate is more prone for attrition (not written south because there the total attrition is also high)
15. Theft cases: central zone max
16. Issues with the manager: South zone and north zone max, then east
17. Better salary : (majorly in north and south) South max, then north
18. Lack of growth: Highest in the south zone, then west.
19. Highest order ranking employee attrition occurs in the "ho" zone only.
20. E1 attrition is highest in South Zone.
21. Employees with M2 and E1 grade are prone to leave the job early.
22. Employees with grade E2 leave the job after a good tenure.
23. M1, M3 and M4 are having comparable tenure.
24. North and west have a comparable and less tenure.
25. South and HO zones have comparatively good tenure with south being better.
26. Females have a better tenure than males.
27. All zones except "ho" zone have a range of 25-30 years for attrition, but for "ho" zone, it is high.
28. E1 has the lowest age range for attrition, then comes e2.
29. M3 and M1 have a comparable range of ages. M2 has a higher range, then comes M4 and CXO are highest referring to the retirement age.
30. MBA Graduates have left the job at more age.
31. North and central have a comparable and good Job satisfaction rate.
32. "ho" zone have comparably less job satisfaction.
33. East, West and South have comparable and average ob satisfaction rate.
34. MBA graduates have better job satisfaction rate than B tech ones, but there are some exceptions
35. M3 has the worst job satisfaction rate, followed by E2
36. E1 and CXO have an average job satisfaction rate
37. M1 and M2 have comparable and good job satisfaction
38. M4 has the best job satisfaction
39. Thus the age and monthly income are correlated to each other with a score of 0.8.
40. Attritions are highest in October this maybe because of the reason that half of the financial year is over so they are eligible for some compensation.
41. Attrition is also high in May, which is just the start of the financial year, so if someone is making a switch to another company, he/she will leave in the start of the financial year only.

Model I (Attrition prediction) results:

For the Attrition prediction, we used IBM dataset. There were some differences in the given and the chosen dataset hence we exclude some features to according to the above analysis. We use the IBM dataset to train the model and original dataset is used to validate and test the model.

Following results have been obtained:

Model Name	Test/Validation Accuracy
KNeighborsClassifier	7.3394%
DecisionTreeClassifier	22.6300%
RandomForestClassifier	20.1835%
GradientBoostingClassifier	30.2752%
LinearDiscriminantAnalysis	84.7095%
lightgbm	Varies as we set the cutoff (50% to 73%)

Model II (Tenure predictions) results:

For the Tenure prediction, we used the same IBM dataset. Attrition feature is not used as it determines the working period(tenure) all remaining features are used for tenure prediction after preprocessing the data. Tenure prediction will help to predict the time or month after which a employees may leave the job. So some important factors can be used to stop employees from leaving the job in future like giving bonus, promotion, salary increment.

Model Name	Test/Validation Accuracy(RMSE)
XGBRegressor	0.373208852699763
RandomForestRegressor	0.31837496703996626
ExtraTreesRegressor	0.18876169255379394
lightgbm	0.13300139337354597
GradientBoostingRegressor	0.3895440296919779
SVR	0.12491605453666697

Value for the industry (Model 1)

Attrition prediction model will predict whether at the present state the employee is prone to attrition or not. It can be a real-time process (month-wise depending on the tenure).

It has a good scope for future practices when an employee is prone to attrition, we can plan for his/her promotion, appraisal, a transfer to a better zone or any such thing which will increase his probability to not undergo attrition.

This will have a great impact on the organisation, the organisation can save important employees which are of great value to the organisation.

Value for the industry (Model 2)

Tenure prediction simply refers to the prediction of the tenure of an individual in the company. The age at which in which he/she is joining the company, position in the organisation, zone etc. can be used to predict the tenure of the employee.

This can be of great value to the company, it can be used to line the appraisals of the employee and giving target projects to him/her. We can closely examine certain cases where the tenure is very less and work on them. If the predicted tenure is less of an employee we can launch a scheme or a bond period for them(this can have a negative impact on the employees join also). Thus tenure prediction is having good scope for future practices.

If improvement of the tenure is not the need, then also this model is useful, the organisation can make an organised scheme for things like appraisals, promotions etc. according to the average tenure for that particular state (or features).