Human Resource Dataset

Leveraging HR data can help inform and improve the strategies used by HR and that is including recruitment, training and development, compensation and even turnover rates. This chosen HR dataset is specifically focusing on the organization turnover rates and what constitute to the consequences.

More on the data collected, the organization HR conducted a survey on their employees and the survey is designed carefully to reveal great deal of information about them. For instance, the HR can tell whether the employee would prioritize family versus working hours or job position vs income. Next is to link to their job satisfaction and here come the main question: are they still working for the company?

In this sense, the researcher can potentially evaluate employee's decision to leave or stay in the company with a predictive model. Further emphasis, the aim of this project is not based on subjective point of view where HR would think "the employee left because the job is not good" but on an objective insights which is reflecting on employee's opinion and their job experience.

Features Description

Age	Attrition	BusinessTravel	DailyRate	Departme	DistanceF	Education	Education	Employee	Employee	Environm	Gender	HourlyRat	JobInvolv	JobLevel	JobRole	JobSatisfa	MaritalSta
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Scien	1	1	2	Female	94	3	2	Sales Exec	4	Single
49	No	Travel_Frequently	279	Research	8	1	Life Scien	1	2	3	Male	61	2	2	Research	2	Married
37	Yes	Travel_Rarely	1373	Research	2	2	Other	1	4	4	Male	92	2	1	Laborator	3	Single
33	No	Travel Frequently	1392	Research	3	4	Life Scien	1	5	4	Female	56	3	1	Research	3	Married

Exploring the dataset, there are a total of 35 columns and close to 1500 rows. The number of columns are consider good which could open up to a lot of possible analysis. The dataset contains employee's background and work-related activities.

Project Brief and Problem Statement

An organization with a bad turnover rate is costly and is affecting the productivity. This is a severe situation where every manager's nightmares to see employees resigning because of the loss of valuable knowledge and experience. On the other hand, HR would experience the frustration to find a suitable replacement every time which involves time, money and risk.

The approach here is to develop a predictive model to predict who will be likely to resign. This allows HR to act quicker to retain the employee rather than to find out the reason after the exit interview. That will be too late!

In addition, the researcher hopes to identify better retention employee strategies through the data so that the HR can consider to adopt.

Predictive Model Workflow

Data Pre-Processing – Handle missing values, remove useless variables, bin variables such as the job satisfaction level. Detect the outliers which is the odd one out and identify them.

Early Data Analysis – Identify 2 groups of employees who stay or had resign. To better understand both group decisions of whether is because of income, working hours, etc

Model Training – Supervised training. Manage imbalance dataset (80% test and 20% accuracy). Using N Naïve Bayes, k-nearest neighbors algorithm and random forest to train and text. Objective here is to form 3 themes to train such as old age, overtime and salary or all three combinations.

Model Validation – Using K-Fold/Confusion Metric techniques to validate how good is the model

Model Predictions – After the best model selected, generate the outcome and do tuning to the best interest of the results

Conclusion and recommendation – Focus on which employees are leaving and why.

7 Questions

The first 3 questions is to understand the data better before performing the predictive modelling.

1) What is the turnover rate in the company (%)
2) Measure the average of employee satisfaction
3) Which factors is having an direct impact to leaving the company (Correlation)
4) Train predictive model with 10 or more features based on supervised learning (Classification)
5) Test with Naïve Bayes, KNN or random forest
6) Which model is the best for the prediction model? Why?
7) Predict employees who are like to leave

Possible Challenges

- Early rectification on outliers (Keep or remove) which might affect the outcome
- High Imbalance data. Eg, More employees stay compared to employees leaving (Stratify technique)
- Models that don't flare well with the features selected – The feature that does not resolute to the outcome.
- Wrong model picked which lead to poor prediction quality

These are preliminary questions which later stage may have more questions raised in the effort of establishing the connections between the variables.