# Employee Attrition – predictive analysis

# **Sowmya Chavali DSC 680 https://github.com/sowmyachavali/DSCPortfolio**

# Which Domain?

This project will deal with employee attrition data in business domain. Hiring and retaining top talent is an extremely challenging task for the HR department, which requires skill, capital, and time. Business owners spend up to 40% of their time on simple tasks not resulting in direct profit, such as hiring. For a new hire, 15-20% of their salary is spent on just recruitment. For better performance of the companies, the HR department has to take measures to reduce employee attrition.

I will attempt to understand the likelihood of an employee leaving the company and identifying the key indicators for the attrition. Post the analysis the company can adopt strategies to improve on employee retention. This is a supervised classification problem where the target is a binary variable, 0 and 1 for active employee and ex-employee.  In this study, our target variable is the probability of an employee leaving the company.

Listing the references -

1. <https://www.indeed.com/career-advice/career-development/reasons-employees-leave> - Gives the top reasons employees decide to quit their jobs.
2. <https://www.flexjobs.com/employer-blog/work-life-balance-vital-employee-retention/> - Having a healthy work-life balance means that employees will be happier when they come to work.
3. <https://www.villanovau.com/resources/hr/importance-of-job-satisfaction-in-the-workplace/> - Respect, trust, healthy environment, Pay and benefits are few factors which may give job satisfaction.
4. <https://www.15five.com/blog/create-a-happy-workplace/> - If the workplace environment is happy and lively then there will be less chances of attrition. Happier the employees are, more successful the company will be.
5. <https://www.cnbc.com/2020/02/20/these-are-the-12-companies-workers-dont-want-to-leave.html> - This article describes what high-retention companies are getting right.
6. <https://www.founderscircle.com/high-startup-turnover-rate/> - Companies invest heavily in recruiting. A bit more focus on retention might pay significant dividends
7. <https://www.sciencedaily.com/releases/2018/07/180705205142.htm> - People are more likely to change jobs when they are younger as they are more open to a new experience.
8. <https://smallbusiness.chron.com/effects-performance-appraisals-employees-17558.html> - Performance appraisals have a number of positive effects on employees and the entire business.
9. <https://www.businessinsider.com/companies-ranked-by-turnover-rates-2013-7>
10. <https://smallbusiness.chron.com/negative-impacts-high-turnover-rate-20269.html> -

High turnover rates can also contribute to lost productivity, employee burnout, and low employee engagement among employees who continue to work for your organization.

# Which Data?

In this case study, the IBM HR dataset was sourced from Kaggle.**[1]**

<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv>

Dataset contains employee data for 1,470 employees with various information about the employees. I will use this dataset to predict when employees are going to quit by understanding the main drivers of employee churn**[1]**.

As stated on the IBM website**[2]**"This is a fictional data set created by IBM data scientists".  
Its main purpose was to demonstrate the IBM Watson Analytics tool for employee attrition.

The dataset contains 1,470 rows and 35 columns.

**Columns**

Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager.

# Research Questions? Benefits? Why analyze these data?

* What is attrition wrt age and gender?
* What's the attrition percentage by job role? Highest and lowest attrition rates based on job role.
* Impact of income towards attrition
* Understanding how Job role and years at company impact an employee to quit the job
* Dependency of business travel in determining attrition
* What is the dependency of years of experience in current role to the target variable?
* Are recently hired managers providing higher job satisfaction to employees?
* What's the working environment by job role?
* How does performance appraisal impact the attrition rate?

An in depth look into the attrition by analyzing these features benefits the company to come up with strategies to improve on the shortcomings for the company growth.

**Process employed-**

* data collection, data cleaning, Exploratory Data Analysis
* feature selection, feature engineering
* model selection, model tuning

# What Method?

This analysis will be done to leverage the power of data science to reduce employee turnover and optimize the HR department. Idea behind this is to develop a model which predicts employees that are more likely to quit.

Modeling is one of the best ways to validate our work. For this, we need to split the datasets into two parts called training set (80%) and testing set (20%). Basically, a training set is used to train the model, and the test set is used to test the model. This is to avoid the overfitting the model to reach high accuracy. After completing data wrangling we split the sample data following ways.

Splitting the train and test sets

X\_train, X\_test, y\_train,y\_test= cross\_validation.train\_test\_split(X,y,test\_size=0.2)

Once we have done the split, we will proceed to train and validate the model. We can use several models to achieve our goal. The models that can be considered in this study are:

**a)Naive Bayes[3]**

Naive Bayes is a simple classification technique that has attracted attention for its simplicity and that all variables are conditionally independent of each other.

For the HR dataset, the Naive Bayes classifier calculates the posterior probability of the target using each of the predictors with the assumption that all the predictors are independent of each other.

**b)Support Vector Machine[4]**

A SVM is a supervised learning algorithm that implements the principles of statistical learning theory and can solve linear as well as nonlinear binary classification problems.

A support vector machine constructs a hyper-plane or set of hyperplanes in higher dimensional space for achieving class separation. The intuition here is that a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class- the larger the margin the lower the generalization error of the classifier.

For the HR Turnover dataset, the support vector uses each of the predictors and constructs a set of hyperplanes in n-D space and chooses the hyperplane that provides the best class separation for the target.

**c)Random Forest** **[5]**

It is a popular ML method that is capable of solving both regression and classification. Random Forest is a brand of Ensemble learning, as it relies on an ensemble of decision trees. It aggregates Classification (or Regression) Trees. A decision tree is composed of a series of decisions that can be used to classify an observation in a dataset.

Random Forest fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. Random Forest can handle a large number of features and is helpful for estimating which of your variables are important in the underlying data being modeled.

The project plan may vary based on further analysis.

# Potential Issues?

There are inherent limitations when fitting machine learning models to smaller datasets.[6] As the training datasets get smaller, the models may have fewer examples to learn from, increasing the risk of overfitting.

# Concluding Remarks

With this project, predictions can be made on employees who are more likely to leave or stay in the company. As the company generates more data on its employees, the algorithm can be re-trained using the additional data and generate more accurate predictions.

**References:**

1. <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv>
2. <https://community.ibm.com/community/user/home>
3. <https://machinelearningmastery.com/naive-bayes-for-machine-learning/>
4. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>
5. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
6. https://www.datarobot.com/blog/using-small-datasets-to-build-models/