HR Attrition Prediction

Thivyadharsine[1], Dennisa Molly[2], Vandhana[3], Aiswarya[4], Sujitha[5], Varsha[6], M.S.Sassirekha[7]
[1,2,3,4,5,6]*Department of Computer Applications, Thiagarajar College of Engineering, Madurai-15*
[7]*Assistant Professor, Department of Computer Applications, Thiagarajar College of Engineering, Madurai-15*

Abstract **- The HR Attrition Case Study is a fictional dataset which aims to identify important factors that might be influential in determining which employee might leave the firm and who may not. In this, we analyzed the dataset Employee Attrition to find the main reasons why employees choose to resign. Firstly, we utilized the correlation matrix to see some features that were not significantly correlated with other attributes and removed them from our dataset. Secondly, we selected important features by exploiting Random Forest, finding monthly income, age, and the number of companies that significantly impacted employee attrition.**

Index Terms **– this loss immediately. Attrition is a major problem in many organizations. Employees leave for personal reasons or move to more promising jobs.**
KNN-K-Nearest Neighbours algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well-suited category by using K- NN algorithm.
Random Forest - Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
Prediction - Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new test data to forecast a particular outcome.
Bagging-It follows a parallel technique (Independent models/Predictions). This gives equal weights to all models
Over fitting - It is another concept under Random Forest Algorithm. Random Forest Algorithm is used to avoid "Overfitting" data.

INTRODUCTION

Employee attrition is defined as the natural process by which employees leave the workforce for example, through resignation for personal reasons or retirement and are not immediately replaced.

Attrition is an inevitable part of any business. There will come a time when an employee wants to leave your company for either personal or professional reasons. But when attrition crosses a particular threshold, it becomes a cause for concern. For example, attrition among minority employee groups could be hurting diversity at your organization. Or, attrition among senior leaders can lead to a significant gap in organizational leadership.

In order for an organization to continually have a higher competitive advantage over its competition, it should make it a duty to minimize employee attrition Therefore, for the better development of corporation, it is essential for the leader of companies to know the main reasons why their employees choose to leave the company, then take relevant measures to improve their company's productivity, overall workflow and business performance.

METHODOLOGY

The methodology used here is the Random Forest Machine Learning Algorithm. Random forest is a supervised learning algorithm. It is based on the concept of ensemble learning and is usually trained with the "bagging" method. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Examining the dataset

The dataset contains 35 different attributes like Age, Business Travel, Daily Rate, Department, Distance , Marital Status, Monthly Income, Number of companies worked, Over18, Over Time, Percent Salary Hike, Performance rating, Relationship Satisfaction, standard working hours, Stock option level, Employee field, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction , Total working years, training

times last year, work-life balance, Years at company , Years in current role, Year since last promotion, years with current manager. There are around 1470 records, each observation corresponding to an employee.

Data Cleaning

Data cleaning is one of the most important parts of data mining. The steps involved in data cleaning may include removal of unwanted observations, fixing structural errors, managing unwanted outliers and handling missing data.

Here the data visualization technique heatmap is used to visualize the attributes which have more attention and matplotlib is used to visualize outliers. The fields with null values are replaced with 0.

Visualizing the data

Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.The matplotlib is used to plot the data.

Data Preprocessing

The process that involves transforming the raw data into a model-able format is called data pre-processing. In the HR dataset there are 4 irrelevant columns namely EmployeeCount, EmployeeNumber, Over18, and StandardHour. Those columns are to be removed for more accuracy.

Next, we have to encode the categorical variables to 1s, and 0s. The variables to be encoded are, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime.

Create dummy variables for more accuracy.

Splitting data to training and testing

training set—a subset to train a model.

test set—a subset to test the trained model.

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

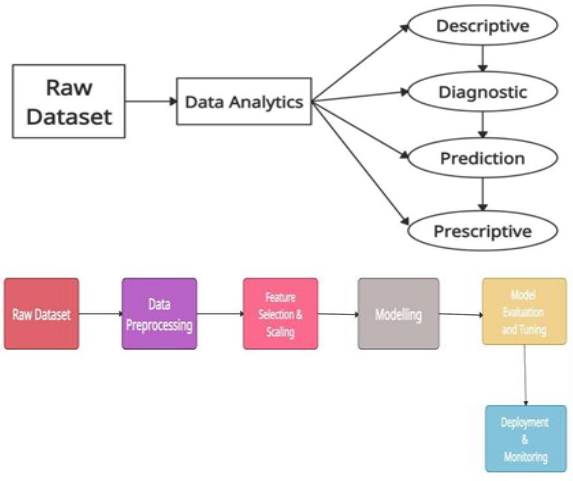
We used train_test_split from the sklearn package to split the data. 70% of the dataset is used for training and 30% is used for testing the data.

Model Execution

To train the model we need to import the model. As we are using the Random Forest classification algorithm we need to import the RandomForestClassifier class from sklearn.linear_model library. An object is created for the RandomForestClassifier class to implement its methods. In this class we have a fit() method whose parameters are the train values. After the model is trained using the training data it is required to predict using the predict() method and test data. And the output is predicted values of test data. According to the Random forest classifier the most important feature for predicting the result is Monthly Income and the least important feature is jobRole_Manager.

Figure and Tables

Prediction Process:

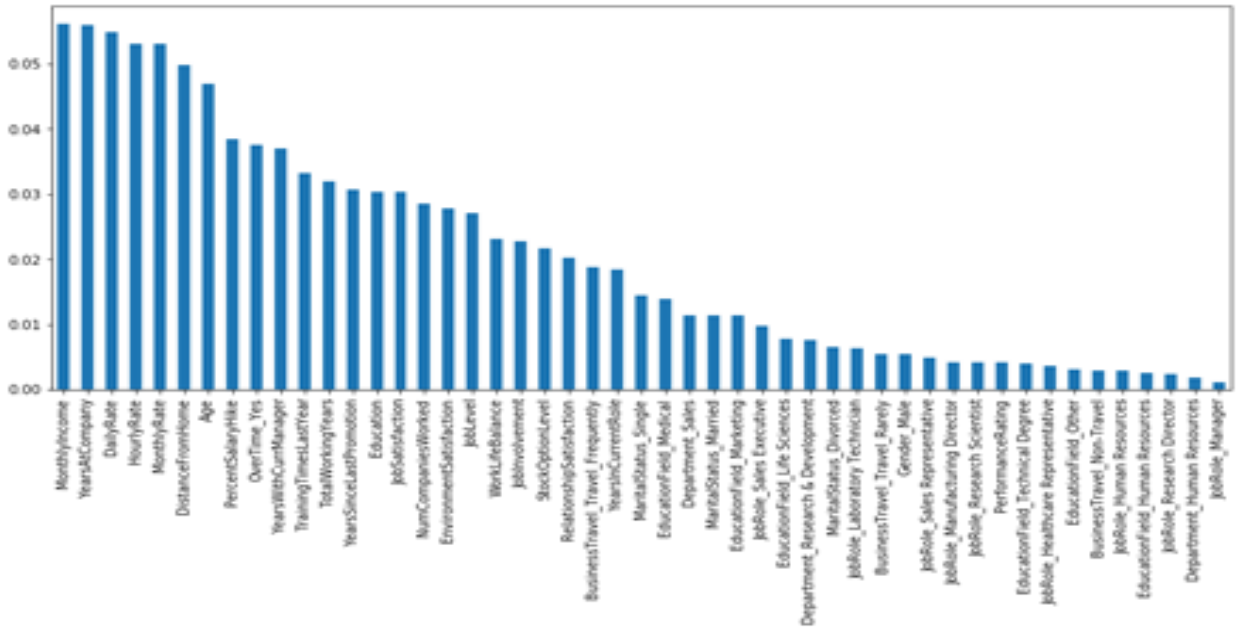


Comparison of Classification Algorithm:

	K-Nearest Neighbours Algorithm	Random Forest Classifier Algorithm
Description	It classifies the data point on how its neighbor is classified.	It is a classification algorithm made up of several decision trees.
Purpose	It is used to solve both classification and regression problems.	It is used to solve regression and classification problems
Performance	It is robust to noisy training data and is effective in case of a large number of training examples.	This algorithm can handle high dimensional spaces as well as a large number of training examples.

Confusion Matrix for Training data	Confusion Matrix: [[915 7] [145 35]]	Confusion Matrix: [[988 0] [19 169]]
Confusion Matrix for Test data	Confusion Matrix: [[299 12] [55 2]]	Confusion Matrix: [[241 4] [39 10]]
Accuracy	81%	85%
	According to the Random Forest classifier the most important feature for predicting the result is Monthly Income and the least important feature is jobRole_Manager.	Using K-Neighbors Classifier for finding the best number of neighbours with the help of misclassification error.

Key features for deciding the result:
pd.Series(rf.feature_importances_,index = X.columns).sort_values(ascending = False).plot(kind = 'bar',figsize = (14,6));



ACKNOWLEDGMENT

We are grateful to our respectable teacher, Mrs. M.S.Sassirekha, for her continuous support and presence whenever needed. We would also like to thank our adjunct faculty Mr. Anbarasan for his advice and contribution to the project and the preparation of this report. Last but not the least, we would like to thank everyone who is involved in the project directly or indirectly.

CONCLUSION

We found the main reasons why employees choose to resign. We analysed that accuracy obtained through Random Forest algorithm having 85% accuracy when compared to the K-Nearest Neighbours algorithm which has 81% accuracy. Hence, the outcome of the algorithms on the same dataset reveals that Random Forest algorithm outperforms than K-Nearest Neighbours algorithm for this particular dataset if accuracy is the metric preferred.

REFERENCE

[1] https://iopscience.iop.org/article/10.1088/1757-899X/830/3/032001/pdf
[2] https://d1wqtxts1xzle7.cloudfront.net/63648492/48120200616-4699-j64310-with-cover-page- v2.pdf?Expires=1635266238&Signature=cX5brvJ6j1Smwf-nQtYSABO~0C-RD2PQKHGT~cDJVqGPeWSmCjdsQKhyLuQSBFwbEKv88-MMORCFee-desJuuI~nCzJbLFuXMKQKa8ffg1BE4 34

PUqeTE05fz-XyfPMn7Fe0tSlTg3Rajqnn2KSm
JWWkQrU5yY7n~N-9hCmr55Goi6nz-OEwPL
DblcdMY9~x2T9xcSX63GuZVcA5HVka3Yq
7Brjqi0O77CRyALSA4P7ry3i5jUIXwqtT3x5~6
DbfOODwJEogfc2uCj7nk-ndhRHDrWzkkFpGp
fwvo4K041nxrBIIoq0-MUOiFenwWmkeixbDU
K65p7HTGsdvBVAYxA&Key-Pair-Id=APKAJ
LOHF5GGSLRBV4ZA

[3] http://citeseerx.ist.psu.edu/viewdoc/download?d
oi=10.1.1.1012.2947&rep=rep1&type=pdf

[4] https://www.researchgate.net/profile/Shawni-
Dutta/publication/341878934_Employee_attritio
n_prediction_using_neural_network_cross_valid
ation_method/links/5ed7becf299bf1c67d352327/
Employee-attrition-prediction-using-neural-
network-cross-validation-method.pdf