Name- Pravin Madhav Bhagwat
Internship Program- Data Science with Machine Learning And Python
Batch- JAN 2022 - MAR 2022
Certificate Code- TCRIB2R232
Date of submission- 5th APRIL 2022

Technical Coding Research Innovation, Navi Mumbai,
Maharashtra, India-410206

HR EMPLOYEE ATTRITION PREDICTION

A Case-Study Submitted for the requirement of
Technical Coding Research Innovation

For the Internship Project work done during

DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM

by
Pravin Madhav Bhagwat (TCRIB2R232)

Rutuja Doiphode
CO-FOUNDER & CEO
TCR innovation.

Name- Pravin Madhav Bhagwat
Internship Program- Data Science with Machine Learning And Python
Batch- JAN 2022 - MAR 2022
Certificate Code- TCRIB2R232
Date of submission- 5th APRIL 2022

HR Employee Attrition Prediction

Pravin Madhav Bhagwat
Department of Computer Engineering, PDEA’S College of Engineering Manjari (Bk), Pune
Sacitribai Phule Pune University Pune, India

Abstract - In this paper, we have analyzed the HR Employee Attrition database to determine the main reasons why employees choose to resign or stay in the organization. Firstly, we will see the correlation matrix and identify the features that are not so correlated with each other and remove them from the dataset. Secondly, by using the Random Forest algorithm we selected some features which has huge impact on the employee attrition like age, salary, distance.

I. INTRODUCTION

The outcome of much research shows that the most valuable asset and important resource in organizations are their employees. Now a day due to increased competition and improved requirements in employees’ proficiency determine the attrition rate. Employee attrition is considered to be a serious issue for organizations. The cost of searching and training employees is very high. Organizations need to search, hire and train new employees. The loss of experienced workers especially high performers is difficult to manage and is negatively related to the success and performance of organizations. The study focuses on the variables that may lead to controlling the attrition rate of the employee. The problem of employee turnover has turned to eminence in organizations because of its pessimistic impacts on issues on workplace self-esteem and efficiency. The organizations deal with this problem is by predicting the risk of attrition of employees using machine learning techniques thus giving organizations to take proactive action for retention.

II. PROPOSED SYSTEM

Initially the data is downloaded is pre-processed first so that we can extract important features like Monthly Income, Last Promotion Year, Salary Hike and etc. that are quite natural for employee attrition. Dependent variables or Predicted variables are the ones that help to get the factors that are mostly dependent on employee-related variables. For example, the employee ID or employee count has nothing to do with the attrition rate.

Exploratory Data Analysis is an initial process of analysis, in which you can 0summarize the characteristics of data to can predict who, and when an employee will terminate the service. The system builds a prediction model by using the random forest technique. It is one of the ensembles learning techniques which consists of several decision trees rather than a single decision tree for classification.

The techniques perform dependent variable analysis and word-formation vector to evaluate the employee churn. Hence, by improving employee assurance and providing a desirable working environment, we can certainly reduce this problem significantly.

III. HR Employee Attrition Dataset

We have provided this dataset in order to help to find a solution to this problem. This dataset contains a total of 35 attributes out of Attrition is the dependent attribute. We have come to a conclusion that with the help of this dataset we might be able to find a solution to this problem. These are the features that are present in our dataset.

Name- Pravin Madhav Bhagwat
Internship Program- Data Science with Machine Learning And Python
Batch- JAN 2022 - MAR 2022
Certificate Code- TCRIB2R232
Date of submission- 5th APRIL 2022

IV. FEATURE SELECTION

Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature to your output variable. Feature importance is an inbuilt class that comes with Tree-Based Classifiers, we will be using Extra Tree Classifier for extracting the top features for the dataset.

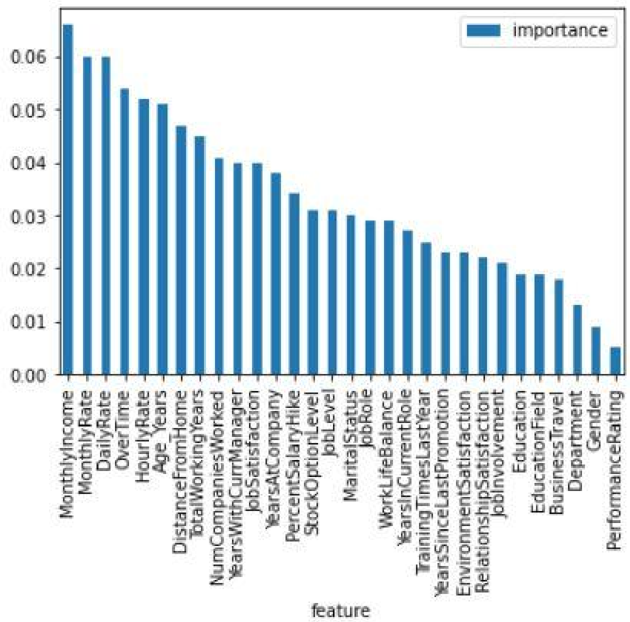


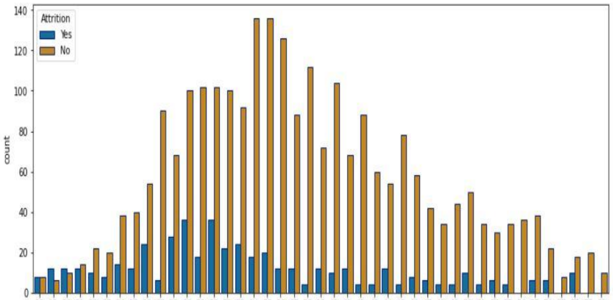
fig- Feature Importance

The diagram above represents the feature importance of each feature of our dataset with the help of this feature importance method we could analyze that the features like Monthly income, Age , Daily rate , Hourly rate, etc are some of the significant attributes. Along with that, we came to the conclusion that the features like Business travel Gender, Department, and Performance rating are having the least impact on our output variable Attrition. Therefore we can neglect these features beforehand. After applying feature selection methods the following attributes are selected for model designing.

V. EXPLORATORY DATA ANALYSIS

In this part we are going to analyze the relationship between the attributes and the output variable. As there are many attributes available to us we cannot show each and every attribute relation. So for simplicity, we are going to use the Age attribute as an example and show the relation.

Attrition Count by using Age:



As shown in the figure above we can clearly see that at the age of 34-36, the rate of attrition is the highest among all ages. At the age of 50 and above the chances of attrition are at the lowest.

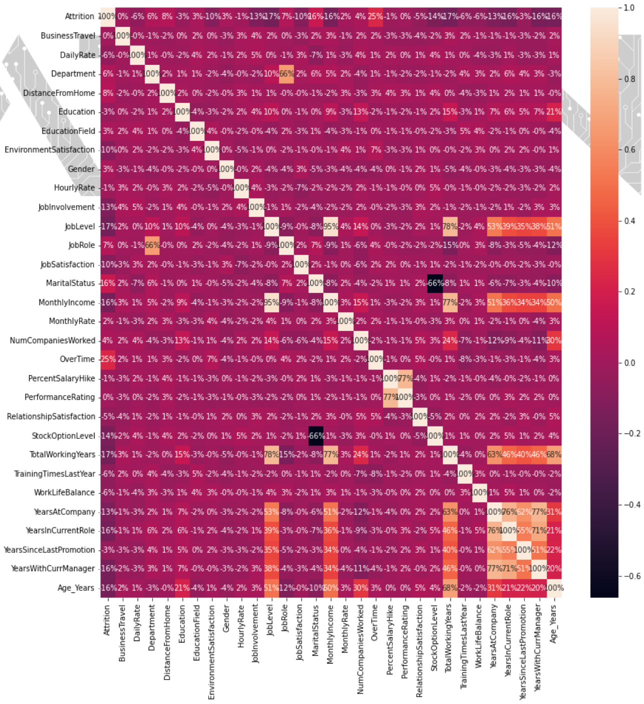


fig- Correlation Map

Name- Pravin Madhav Bhagwat
Internship Program- Data Science with Machine Learning And Python
Batch- JAN 2022 - MAR 2022
Certificate Code- TCRIB2R232
Date of submission- 5th APRIL 2022

VI. IMBALANCED DATASET

In the dataset 10% of records are labeled with class, YES, and the remaining 90% of records are labeled with class NO. This type of dataset is called an imbalanced dataset and can have an adverse effect on the performance of the model it makes the model biased towards the majority class of output variables. Therefore handling an imbalanced dataset becomes a necessary task for this type of problem statement.

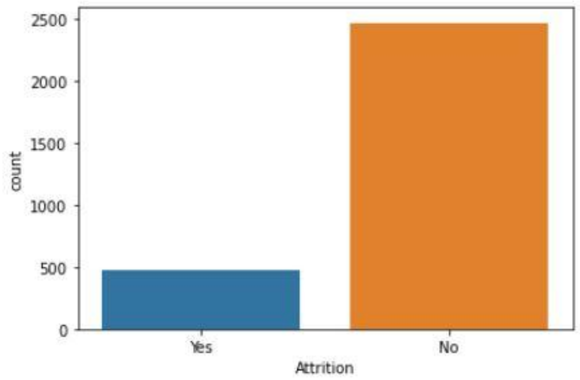
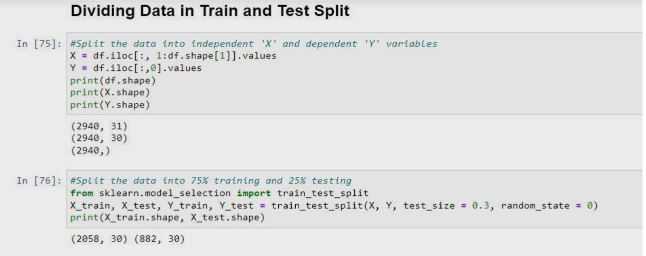


fig -Imbalanced Data

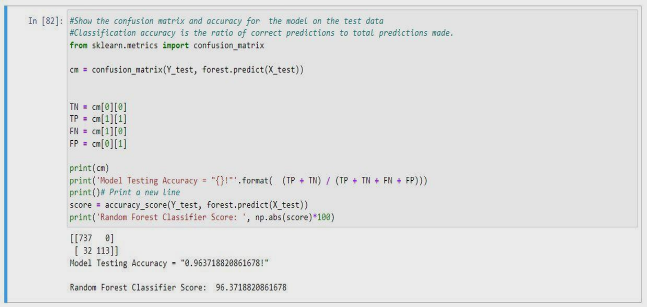
TRAINING & PREDICTION

Now after the analyzation of the dataset, we can build a model by application of the Random Forest Classification algorithm to train the machine learning model because this algorithm can work effectively with a large number of features. I split the dataset as 75 % for training and 25 % for testing. Below is the implementation of training and predicting the machine learning model to achieve the aim.

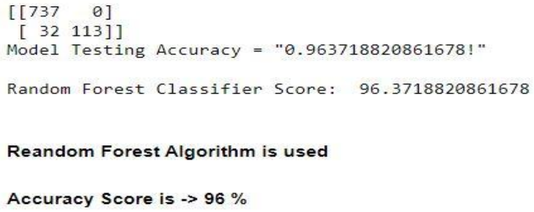


Dividing Data in Train and Test Split

```
In [75]: #Split the data into independent 'X' and dependent 'Y' variables
         X = df.iloc[:, 1:df.shape[1]].values
         Y = df.iloc[:,0].values
         print(df.shape)
         print(X.shape)
         print(Y.shape)

         (2940, 31)
         (2940, 30)
         (2940,)

In [76]: #Split the data into 75% training and 25% testing
         from sklearn.model_selection import train_test_split
         X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3, random_state = 0)
         print(X_train.shape, X_test.shape)

         (2058, 30) (882, 30)
```



```
In [82]: #Show the confusion matrix and accuracy for  the model on the test data
         #Classification accuracy is the ratio of correct predictions to total predictions made.
         from sklearn.metrics import confusion_matrix

         cm = confusion_matrix(Y_test, forest.predict(X_test))


         TN = cm[0][0]
         TP = cm[1][1]
         FN = cm[1][0]
         FP = cm[0][1]

         print(cm)
         print('Model Testing Accuracy = "{}!"'.format(  (TP + TN) / (TP + TN + FN + FP)))
         print()# Print a new line
         score = accuracy_score(Y_test, forest.predict(X_test))
         print('Random Forest Classifier Score: ', np.abs(score)*100)

         [[737   0]
          [ 32 113]]
         Model Testing Accuracy = "0.963718820861678!"

         Random Forest Classifier Score:  96.3718820861678
```



```
[[737   0]
 [ 32 113]]
Model Testing Accuracy = "0.963718820861678!"

Random Forest Classifier Score:  96.3718820861678
```

Reandom Forest Algorithm is used

Accuracy Score is -> 96 %

VII. CONCLUSION

After applying the Random Forest Classification algorithm, the machine learning model will be able to predict employee attrition with an accuracy of 96.37%. This is not the only method to train the model for the prediction of employee attrition. Using various other algorithms, it can be possible to predict but I found that this algorithm works better than all other classification algorithms in the "HR EMPLOYEE ATTRITION DATASET". `

IX. REFERENCES

1) https://www.researchgate.net/publication/351911311_Prediction_of_Employee_Attrition_Using_Machine_Learning_and_Ensemble_Methods
2) S. Jahan, “Human Resources Information System (HRIS): A Theoretical Perspective”, Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129, 2014.
3) https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/