IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12114

Data Science to Analyse Employee Data

Jyothi M B[2], Dhruva L[2]

BE, Department of ECE, DBIT, Bangalore, India[1]

BE, Department of CSE, Dr. AIT, Bangalore, India[2]

Abstract: Data analytics is the process of analysing unprocessed data to derive conclusions. In this paper, we'll examine the trends in employee turnover and the factors that influence them. A model that can forecast whether a specific employee will leave the organisation or not will be developed. The objective is to develop or enhance various retention methods for selected staff. The paper presents data pre-processing, the initial step in data analytics. Techniques for data pre-processing transform unusable data into useful forms. Real-world data are frequently insufficient, inconsistent, and full of inaccuracies. Analysis and prediction are of higher quality when these factors are eliminated. Inference, or the process of drawing conclusions, is the main emphasis of data analytics. In this paper 2 out of top 3 strategies affecting employee turnover are being analysed and graphs plotted. The 3 top features include evaluation v/s exit, average monthly income v/s exit and satisfaction v/s exit.

Keywords: examination of raw or crude data and drawing conclusions- data analytics, employee turnover pattern, data pre-processing, evaluation v/s exit, average monthly income v/s exit and satisfaction v/s exit.

I. INTRODUCTION

In this information era, huge amount of data is being stored, exchanged and conditioned. The volume of data that one has to deal with has exploded to unimaginable levels. Most of the data exists in its crude form and needs to be converted to useful format before analysis. This process of converting raw data into useful format is called data pre-processing. Real world data is [1-6]

- Incomplete: consists of missing attribute values or consists of only aggregate data.
- Noisy: containing errors or outliers.
- Inconsistent: containing discrepancies in code.

II. PROBLEM STATEMENT

In this paper we consider a company doing business in this 21st century world. This company has employed large number of employees working in different departments. The size of the company is huge and has several departments. The company is existent in business from a very long period of time and several employees have already left the company. The company's historical data is well tabulated and records maintained. The company wants to understand what factors contributed most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees. [7-8]

```
In [3]:
import pandas as pd
import numpy as np
import seaborn as sns
%matplotlib inline
```

Figure 1 shows the Python code to import libraries.

III. METHODOLOGY

A. Importing Libraries

Figure 1 shows the Python code to import libraries. We have used three libraries [2]

- NumPy is the fundamental package for scientific computing with Python.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12114

- Pandas is for data manipulation and analysis. Panadas is an open source, BSD- licenced library providing easy-to-use data structures and data analysis tools.
- Matplotlib is a python 2D plotting library. It can be used in Python scripts, Jupyter notebook, web application servers and IPython shells.
- Seaborn is a Python data visualization library based on matplotlib for attractive and informative statistical graphics.

B. Importing data

Figure 2 shows the Python code to import data from respective directory/ file. The data stored in CSV format is being imported. [3]

C. Checking for missing values [4]

It is very essential in data pre-processing to check for missing values. Figure 3 shows the Python code to check for missing values. In this attempt no missing values were found.

D. Renaming and rearranging the columns

It is essential to rename the columns so that analysis is effective. Figure 4 shows the process of renaming the columns and figure 5 shows an effort to move the column ‘exit’ to the end as it has to be predicted.

E. Exit rate[5-9]

Exit rate of the employees need to be checked. Figure 6 shows the exit ratio calculation. 76% of the employees stayed and 24% of employees exited.

```
In [4]:
#Importing Data
df = pd.read_csv(r'C:\Users\manasaav\datascience\employee-turnover-analysis\data.csv')
```

Figure 2 shows the Python code to import data from respective directory/ file.

```
In [5]:
#Checking whether our data contains any missing value or not
df.isnull().any()

Out[5]:
satisfaction_level       False
last_evaluation          False
number_project           False
average_montly_hours     False
time_spend_company       False
Work_accident            False
left                     False
promotion_last_5years    False
sales                    False
salary                   False
dtype: bool
```

Figure 3 shows the Python code to check for missing values.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12114

IV. EVALUATION V/S EXIT

- There is a bimodal distribution for those that had an exit.
- Employees with low performance tend to leave the company more.
- Employees with high performance tend to leave the company more.
- The sweet spot for employees that stayed is within 0.6-0.8 evaluation. Figure 7 shows the employee monthly hour distribution.[9-12]

V. AVERAGE MONTHLY HOURS V/S EXIT

- Another bimodal distribution for employees that exited.
- Employees who had less than 150 hours of work left the company more.
- Employees who had more than 250 hours of work left the company more. Figure 8 shows the employee evaluation distribution.

```
In [7]:
#Renaming the columns
df = df.rename(columns={'satisfaction_level': 'Satisfaction',
                        'last_evaluation': 'Evaluation',
                        'number_project': 'ProjectCount',
                        'average_montly_hours': 'AverageMonthlyHours',
                        'time_spend_company': 'YearsAtCompany',
                        'Work_accident': 'WorkAccident',
                        'promotion_last_5years': 'Promotion',
                        'sales' : 'Department',
                        'left' : 'Exit'
                        })
df.head()

Out[7]:
```

Figure 4 shows the process of renaming the columns

```
In [6]:
#Moving the column 'Exit' to the end which is to be predicted
front = df['Exit']
df.insert(0, 'Exit', front)
df.head()

Out[6]:
```

Figure 5 shows an effort to move the column ‘exit’ to the end as it has to be predicted.

```
Exit_Rate = df.Exit.value_counts()/len(df)
Exit_Rate

Out[9]:

0    0.761917
1    0.238083
Name: Exit, dtype: float64
```

Figure 6 shows the exit ratio calculation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12114

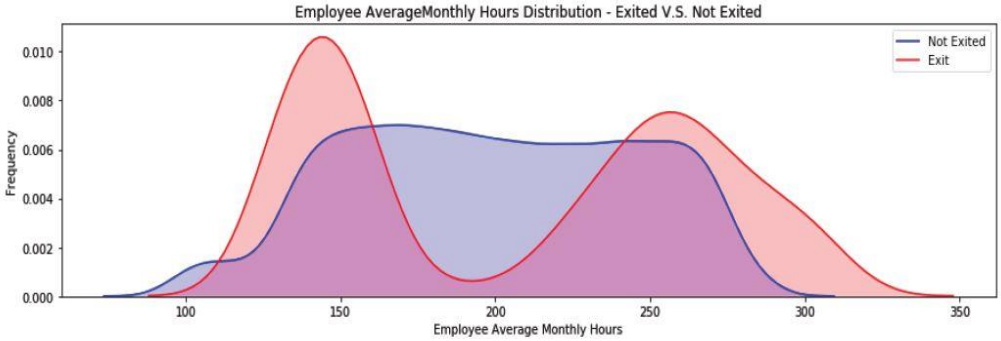

Figure 7 shows the plot of employee average monthly hours distribution.

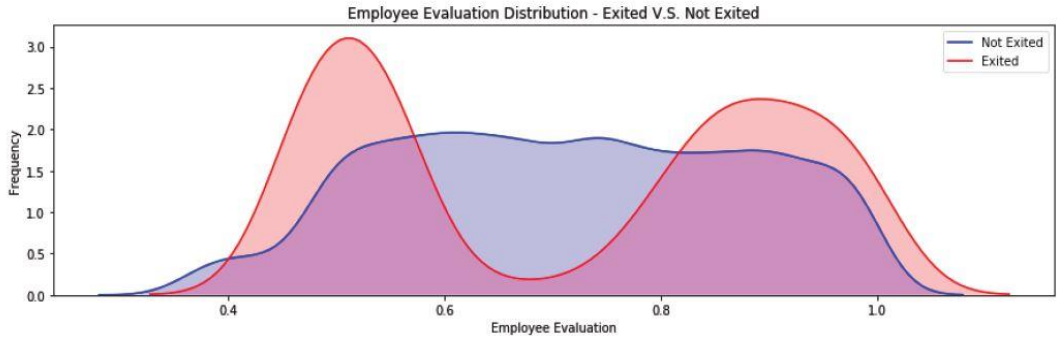

Figure 8 shows the employee evaluation distribution.

VI. CONCLUSIONS

A company proactive in business in this 21st century world had many workers leaving the company. Data analytics had to be carried out on the data –both historical and present trend to draw inference. The goal was to create or improve different retention strategies on targeted employees working in different departments of the company. A python code was written and executed in the Jupyter platform to analyse and draw conclusions. The first step in data analytics- data pre-processing was successfully carried out and exit ratio calculated. 2 out of top 3 strategies affecting employee turnover are being analysed and graphs plotted. The 3 top features include evaluation v/s exit, average monthly income v/s exit and satisfaction v/s exit.

REFERENCES

[1] Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. algorithms, 4(5), C5.

[2] W. C. Hong, S. Y. Wei, and Y. F. Chen, “A comparative test of two employee turnover prediction models”, International Journal of Management, 24(4), 808, 2007.

[3] Gao, Ying. "using decision tree to analyze the turnover of employees." (2017).

[4] Mauricio A. Valle & Gonzalo A. Ruz (2015) Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms, Applied Artificial Intelligence, 29:9, 923-942, DOI: 10.1080/08839513.2015.1082282.

[5] Metz, Charles E. "Basic principles of ROC analysis." Seminars in nuclear medicine. Vol. 8. No. 3. WB Saunders, 1978.

[6] Lessmann, Stefan, and Stefan Voß. "A reference model for customercentric data mining with support vector machines." European Journal of Operational Research 199.2 (2009): 520-530.

[7] T. Fawcett, “An introduction to ROC analysis”, Pattern Recognition Letters 27 (8), 861–874, 2006.

[8] Raschka, S. (2015). Python machine learning. Packt Publishing Ltd.

[9] Morgan, J.N., Sonquist, J.A.: Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. 58, 415–434 (1963).

[10] Efron, B.S., Hastie, T.: Computer Age Statistical Inference. Cambridge University Press, Cambridge (2016).

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023
DOI: 10.17148/IJARCCE.2023.12114

[11] Géron, A.: Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media (2017).
[12] Zhao, Yue, et al. "Employee turnover prediction with machine learning: A reliable approach." Proceedings of SAI intelligent systems conference. Springer, Cham, 2018.

RESEARCH GUIDE

VISHESH S born on 13th June 1992, hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He also worked as an intern under Dr. Shivananju BN, former Research Scholar, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication, BAN and Medical Electronics. He is also the Founder and Managing Director of the corporate company Konigtronics Private Limited. He has guided over a hundred students/interns/professionals in their research work and projects. He is also the co-author of many International Research Papers. He is currently pursuing his MBA in e-Business and PG Diploma in International Business. Presently Konigtronics Private Limited has extended its services in the field of Software Engineering and Webpage Designing. Konigtronics also conducts technical and non-technical workshops on various topics.