EMPLOYEE ATTRITION PREDICTION

A PROJECT REPORT

Submitted by

AATHMIHAN S B (210701006)

AKSHITHAA B K (210701024)

in partial fulfilment for the course

CS19643 – FOUNDATIONS OF MACHINE LEARNING

for the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

RAJALAKSHMI NAGAR

THANDALAM

CHENNAI – 602 105

MAY 2023

RAJALAKSHMI ENGINEERING COLLEGE

CHENNAI - 602105

BONAFIDE CERTIFICATE

Certified that this project report **"EMPLOYEE ATTRITION PREDICTION"** is the bonafide work of **"AATHMIHAN S B (210701006)** , **AKSHITHAA B K (210701024)"** who carried out the project work for the subject CS19643 – Foundations of Machine Learning under my supervision.

Dr. P. Kumar	Dr. S. Vinodkumar
HEAD OF THE DEPARTMENT	**SUPERVISOR**
Professor and Head	Professor
Department of	Department of
Computer Science and Engineering	Computer Science and Engineering
Rajalakshmi Engineering College	Rajalakshmi Engineering College
Rajalakshmi Nagar	Rajalakshmi Nagar
Thandalam	Thandalam
Chennai - 602105	Chennai - 602105

Submitted to Project and Viva Voce Examination for the subject CS19643 – Foundations of Machine Learning held on _____.

ABSTRACT

Employee attrition, the phenomenon of employees voluntarily leaving an organization, presents significant challenges to businesses, including loss of talent, decreased productivity, and increased recruitment costs. Predicting employee attrition can help organizations proactively address retention strategies and mitigate its adverse effects. In this project, we propose a machine learning approach utilizing the Random Forest algorithm for predicting employee attrition. Random Forest, a powerful ensemble learning technique, leverages the collective wisdom of multiple decision trees to make accurate predictions. Our model integrates various employee-related features such as job satisfaction, salary, tenure, performance ratings, and demographic information to predict the likelihood of attrition. Through extensive experimentation and validation on real-world employee datasets, we demonstrate the effectiveness of our approach in accurately identifying individuals at risk of attrition. The proposed model not only provides valuable insights into the factors influencing attrition but also empowers organizations to proactively implement targeted retention strategies, thereby fostering a more stable and engaged workforce.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Thiru. S.Meganathan, B.E., F.I.E.,** our Vice Chairman **Mr. M.Abhay Shankar, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) Thangam Meganathan, M.A., M.Phil., Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N.Murugesan, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P.Kumar, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We are very glad to thank our Project Coordinator, **Dr. S.Vinodkumar, M.E., Ph.D.,** Professor, Department of Computer Science and Engineering for their useful tips during our review to build our project.

AATHMIHAN S B (210701006)

AKSHITHAA B K (210701024)

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Employee attrition, a pervasive challenge in modern workplaces, presents a complex interplay of factors that impact organizational stability, productivity, and overall success. As employees voluntarily depart from organizations, they leave behind voids in expertise, disrupt team dynamics, and trigger costly recruitment and training processes. Predicting and preemptively addressing employee attrition has thus become a critical endeavor for businesses aiming to maintain a sustainable workforce and competitive edge in the market. In this project, we delve into the realm of machine learning, particularly focusing on the Random Forest algorithm, to forecast employee attrition with precision and efficacy.

By harnessing the power of machine learning, we seek to construct a predictive model that can discern patterns within diverse sets of employee-related data. This data encompasses not only traditional metrics such as job satisfaction, salary, and tenure but also more nuanced indicators including performance ratings, career advancement trajectories, and demographic profiles. Through the amalgamation of these multifaceted features, our objective is to develop a robust framework capable of accurately identifying employees at risk of attrition.

This project endeavors to go beyond mere prediction; it aims to unravel the intricate web of factors influencing attrition within organizations. By uncovering correlations and insights buried within the data, we aspire to equip businesses with actionable intelligence to proactively address attrition drivers and implement tailored retention strategies. Through a

comprehensive analysis of historical data and iterative model refinement, we strive to provide decision-makers with a powerful toolset for fostering a resilient, engaged, and fulfilled workforce.

Ultimately, this project aspires to transcend the boundaries of conventional attrition management approaches by leveraging cutting-edge machine learning methodologies. By harnessing the predictive capabilities of Random Forest and the wealth of employee-related data at our disposal, we aim to empower organizations to navigate the challenges of attrition with foresight, agility, and strategic precision.

1.2 OBJECTIVE

The objective of this project is to develop a robust machine learning model, utilizing the Random Forest algorithm, to accurately predict employee attrition within organizations. By incorporating a diverse range of employee-related features, such as job satisfaction, salary, tenure, performance ratings, and demographic information, the aim is to construct a predictive framework capable of identifying individuals at risk of voluntary departure. Additionally, the project seeks to uncover underlying patterns and correlations within the data to provide actionable insights for implementing targeted retention strategies, empower organizations to make data-driven decisions, enhance understanding of attrition dynamics, and validate the effectiveness of the Random Forest algorithm in handling complex attrition prediction tasks. Through achieving these objectives, the project aims to assist businesses in proactively addressing attrition challenges and fostering a resilient and engaged workforce.

1.3 EXISTING SYSTEM

In the existing system, HR departments typically rely on traditional methods to identify and address employee attrition. These methods often involve manual processes and subjective assessments, which may lack the predictive power needed to effectively mitigate turnover. HR professionals may conduct exit interviews or administer employee surveys to gather feedback from departing employees, but these approaches are reactive in nature and do not provide early warnings of potential attrition.

Moreover, historical turnover data analysis is often used to identify patterns or trends in employee attrition. While this approach can offer some insights into the factors contributing to turnover, it may not capture the complex interactions between various factors influencing attrition. Additionally, the reliance on historical data means that interventions are implemented after an employee has already decided to leave, limiting the effectiveness of retention efforts.

Another limitation of the existing system is the inability to leverage advanced analytical techniques to predict employee attrition. Traditional methods may overlook important predictors of turnover or fail to account for the dynamic nature of employee engagement and satisfaction. As a result, organizations may struggle to identify at-risk employees early enough to intervene and prevent turnover effectively.

Overall, the existing system for employee attrition prediction relies heavily on manual processes and retrospective analysis, which may not be sufficient to address the challenges posed by turnover. To overcome these limitations, there is a need for a more proactive and data-driven approach to attrition

prediction that leverages advanced analytical techniques and machine learning algorithms.

1.4 PROPOSED SYSTEM

The proposed system aims to revolutionize employee attrition prediction by leveraging machine learning techniques, specifically the Random Forest algorithm. Unlike the existing system, which relies on manual processes and retrospective analysis, the proposed system offers a proactive approach to identifying and addressing employee attrition.

The key innovation of the proposed system lies in its ability to analyze a diverse set of features to predict employee attrition accurately. By considering factors such as demographic information, job-related attributes, performance metrics, and engagement indicators, the model can capture complex relationships and patterns that influence turnover. This holistic approach enables HR departments to identify at-risk employees more efficiently and accurately, thereby enabling proactive intervention to prevent attrition.

The proposed system automates the process of attrition prediction, allowing organizations to analyze large volumes of data and identify potential attrition risks in real-time. By providing early warnings of impending turnover, the system enables HR professionals to implement targeted retention strategies and interventions to address underlying issues and improve employee satisfaction and engagement.

Furthermore, the proposed system offers greater flexibility and scalability compared to traditional methods, allowing organizations to adapt to changing workforce dynamics and evolving business needs. By integrating

data-driven insights into HR decision-making processes, the system empowers organizations to make more informed decisions about talent management and retention strategies.

Overall, the proposed system represents a significant advancement over the existing manual methods for employee attrition prediction. By harnessing the power of machine learning and advanced analytics, organizations can gain deeper insights into the factors driving attrition and take proactive measures to retain their top talent and foster a positive work culture.

CHAPTER 2
LITERATURE REVIEW

Denver and McMahon (1992) define labour turnover as "the movement of people into and out of employment within an organization" while Mobley (1982) defines turnover as "voluntary cessation of membership in an organization by an individual who receives monetary compensation for participating in that organization". Forbes (1971) states that labour turnover means separation from an organization and included promotion, transfer or any other internal movement within the institution. Meaghan et al. (2002) draw attention to controlling attrition; he states that the value of employees to an organization is a very crucial element in the success of the organization. He further states that this value is intangible and cannot easily be replicated, therefore, the managers should control attrition, Mobley (1977) suggests a measure to predict attrition, he says that tenure of an employee is one of the best measures that can be used to predict turnover. Firth et al (2007) try to find out the causes of attrition, he says that there is a range of factors that lead to job-related stress, lack of commitment towards the organization and job dissatisfaction which cause employees to quit. Griffeth et al. (2000) conclude that pay and pay-related variables have a significant effect on employee turnover. Hom & Griffeth (1995) state that several investigations in the past have revealed that organizational commitment and job satisfaction are crucial factors that influence turnover intention. Wanous (1992) focuses on new employee attrition and says that new employees often leave the organization because their expectations are not met which results in a violation of their psychological contract resulting in a turnover. Abassi et al (2000) conclude that there are other factors like inefficient and poor recruitment practices, style of management, lack of recognition, workplace conditions, and a lack of competitive compensation system that cause employees to quit the organization. Louis (1980)

states that attrition takes place because new employees compare their actual experience with their past work experiences. Past work experience plays a significant role in taking the decision to quit in case the new worker‟s expectations are not met. Ongori (2007) focuses on stress as a cause of attrition; he says that the good workers in an organization may tend to leave when they start experiencing signs of occupational stress. This turnover affects the organization adversely in increasing the recruitment and selection costs of the organization.

2.1 KINDS OF ATTRITION

A. Voluntary attrition

Voluntary attrition takes place when the employee leaves the organization by their own will. Pull factors like higher emoluments elsewhere, better opportunities for growth and promotion etc. are responsible for this kind of attrition.

B. Involuntary attrition

Involuntary attrition takes place when the employees leave the organizations due to some negative forces or push factors like faulty promotion policy, biased performance appraisal etc.

C. Compulsory attrition

It takes place due to the rules and regulations of the government and that of the organization as well. It includes attrition taking place due to attaining the age of retirement, completion of tenure etc.

D. Natural attrition

It takes place due to the causes and factors that are beyond the control of the individual and organization as well. These factors may include end of life, insanity etc.

2.2 CAUSES OF ATTRITION

A. Internal causes

These causes are pertaining to the internal environment of an organization. Therefore, they are controllable.

1) Salary

- Insufficient salary
- Delay in payment
- No / delayed increment
- Wage compression

2) Promotion

- Biased promotion
- No / delayed promotion

3) Transfer

- Forceful transfer
- Transfer to a placed employee is not willing to go

4) Workplace Infrastructure & amenities

- lack of hygiene
- lack of basic facilities like water, canteen, etc.

5) Task

- Monotony of task
- Task – labour mismatch

- Team issues
- Lesser job autonomy

B. Instability in leadership

Leading to confusion related to directions and commands which generate frustration among the workforce.

C. Lack of Flexibility

- Lack of flexibility in timing, choice of task etc.
- Introduction of new technology and employee's incompetency/ unwillingness to learn and understand.

D. Lack of job security

- Fear of being expelled/ retrenched/terminated
- Faulty performance appraisal
- Underestimation of performance
- Power distance & politics
- The communication gap between management and workforce

E. External causes

These are the causes which are beyond the control of an organization as they belong to the external environment. These causes may be related to,

- better pay
- chances of promotion
- better perks and
- more fringe benefits in other organizations

F. Individual/Personal causes

- end of life
- marriage
- pregnancy
- shift of family
- mental imbalance
- over – sensitivity
- wish to go abroad
- attrition of the group members
- Self-employment
- Education

2.3 EFFECTS OF ATTRITION

A. Effect on employer/ Organization

- Loss of productivity
- Loss of quality

B. Increase in cost

Attrition results in an increase in costs. These costs may be related to the cost of the exit interview.

1) Cost of staffing

   Cost of travelling allowance, refreshment, experts, placement companies.

2) Cost of Training

Cost of trainers, cost of training equipment and materials, cost of refreshment, cost of technology.

3) Cost of administrative proceedings

   Cost of issuing I – cards, access cards.

4) Cost of signing bonus

   It is given to the works for joining the organization; it is also a significant part of the cost.

   - Loss of consumers and decrease in brand loyalty
   - Loss of goodwill
   - Loss of secrecy in case the key employees leave the organization
   - Loss of key – personnel
   - Lack of competitiveness

2.4 EFFECT ON EMPLOYEE

- Stress from a new job
- Monetary loss
- Effect on career
- Effect on family life
- Loss of skill- if the gap between quitting from one organization to other is long
- Emotional loss, if the bonding with the staff of the previous organization was good

However, it is also possible that the employee gets a better environment and remuneration in the new organization and the things can get positive for him.

Can Attrition Have Positive Effect On the Organization?

Attrition is not always negative; it may have some positive results also. Some of the positive results may include the following,

1. Advantages of new knowledge: New employees bring new knowledge; their knowledge and skill may open new avenues for the organization.

2. Advantage of new technology: It will decrease the cost, thus the price of the final goods or service will be cheaper; further leading to an increase in demand and profits.

3. Introduction of new ideas: New ideas may help in increasing product line and product mix or they may become helpful in starting new joint – ventures and working in collaboration.

4. The lesser negative impact of groupism: Sometimes the existing groups may be rigid or the group members may be reluctant towards others, in such case attrition of a group member may be positive for the organization.

5. Reduction in surplus staff: It will lead to a reduction in the cost of maintaining the surplus employees ultimately leading to the total cost.

6. Chances of bringing in creativity & innovation: New workers may introduce a new style of working, they can have their own methods and they may think differently, all this will promote creativity and innovation in the organization.

7. Creation of a healthy and competitive environment in the organization: The new workforce may be more competitive, old employees may learn from them. They may get inspired and compete with them.

8. Measures to Control Attrition / Retention Strategies Corporate Social Responsibility (CSR) towards employees: It comprises a wide range of intrinsic and extrinsic rewards and motivation. It is concerned with a humanitarian aspect towards the employees of the organization. It is the first and foremost responsibility of an organization to take care of its employees‟ physical and mental wellbeing. CSR towards employees encompasses all monetary and non- monetary aspects. A monetary aspect includes reasonable remuneration, bonus, increment, HRA, post-retirement pension, etc. while the non- monetary aspect may include the congenial environment, fair performance appraisal, recreational activities, learning and development. Both these aspects are equally important while considering control on attrition.

9. Herzberg's Two Factor Theory, CSR towards Employees & Attrition: Hertzberg Two-factor theory describes two factors (Herzberg, Fredrick 1968).

A. Motivators

These factors are related to the intrinsic aspect of the job itself, such as recognition, achievement, personal growth etc.

B. Hygiene factors

These factors are related to the extrinsic aspect of the job such as salary, fringe benefits, work conditions, status, job security etc.

Hygiene factors are essentials, they do not show a direct contribution to productivity but their absence certainly leads to a decrease in production.

Motivators have a positive correlation with productivity; their presence results in an increase in productivity and their absence leads to a fall in the same. Thus both these factors should be paid attention to boost the morale of the workers leading to lesser attrition as morale and attrition have an inverse relationship i.e. Higher the morale, lesser will be the attrition and vice-versa

Applying Emotional Intelligence: Emotional intelligence refers to the ability and capacity to know and control own emotions and that of others in such a manner that the energies and potentials may be channelized in a positive direction and utilized to enhance productivity. To develop emotional intelligence one has to develop empathy and farsightedness. Following are some ways to apply emotional intelligence to control attrition

- Being proactive
- Lessening communication gap between management and workers
- Devising and communicating career and growth opportunities
- Using intrinsic motivation
- Understanding group dynamics
- Conducting motivational sessions for the employees
- Praising the employee publicly but criticizing privately
- Developing a rapport with the workers

Change in leadership style: Leadership can play a significant role in controlling attrition. With the change in organizational dynamics, the style of leadership should also change. One of the much-lauded styles is transformational leadership. Bass & Avolio (1993) state that transformational leadership comprises of the four dimensions: idealized

influence, inspirational motivation, intellectual stimulation, and individual consideration. Such leadership helps the employees in finding out their hidden talent and latent skills. They come to know about their strengths and the scope to enhance them. This acts as an undercurrent in unleashing their energies with full faith in their capabilities resulting in a passion for work, greater connectivity with the organization and its goals and control on the tendency of the workers to leave the organization.

Holistic leadership inculcates a natural sensitivity and empathy towards employees which will significantly increase the belongingness of employees towards the organization. Such leadership will infuse an environment of care and sympathy in organizational relations and ease in working. This will further lead to a decrease in attrition.

Goleman (2001) states about six leadership styles; they are commanding, visionary, affiliative, democratic, pacesetting and coaching leadership style. Out of these six, affiliative leadership, according to Goleman, creates harmony in relations and builds emotional bonds while democratic leadership promotes employees participation in decision making. Both these styles boost the relatedness, belongingness and cohesion in relations which is helpful in decreasing attrition.

Leaders should recognize, promote and praise hard work; employees should be given due credit and compliments. Leaders should be open to discussions and have a welcoming attitude towards the suggestions of the workers. First, they should understand and accept the value of employees and then make the employees feel that they are valuable to the organization; this will bring more openness, harmony, trust in relations. All these factors will be helpful to control attrition.

Flexibility: Flexibility is necessary for a greater degree of coordination, ease and smoothness in the organizational working. It is the demand of

time as in the present context it has become very difficult to manage talent. Undue strictness and rigidity are no more considered the obvious right of the employer. Flexibility can be related to the following factors,

- Time
- Choice of task
- Transfer
- Targets
- Leaves
- Methods
- Place of work in the organization
- Number of breaks

Conducting a stress interview: Exit interviews become instrumental in assessing the level of satisfaction or dissatisfaction of the employee. It should be well planned and questions should be well-f framed. It should focus on the issues like,

- work environment
- Organizational culture
- Peer group
- Senior- subordinate relationship
- Performance appraisal
- Individual growth

In this regard, the following factors should be taken care of

- Questions should be open-ended

- Utmost confidentiality should be maintained
- The process should not be lengthy

Other measures

- Workers‟ participation in management
- Profit-sharing
- Gainsharing
- Fair performance appraisal
- Realistic goals
- Defining career path and demystifying career growth proper succession planning
- Effective communication system

CHAPTER 3

PROJECT DESCRIPTION

3.1 MODULES

3.1.1 DATA COLLECTION

This module serves as the foundation of the project by collecting relevant data from diverse sources such as HR databases, employee surveys, performance evaluations, and organizational records. Data collection involves extracting both structured (e.g., numerical data, categorical data) and unstructured (e.g., text data from surveys, comments) data. The collected data may include information such as employee demographics (age, gender, education), job-related attributes (job role, department, salary), performance metrics (productivity, performance ratings), and engagement indicators (satisfaction scores, feedback). Data cleaning and preprocessing techniques are applied to ensure data quality and consistency, including handling missing values, removing duplicates, and standardizing data formats.

3.1.2 FEATURE ENGINEERING

Feature engineering is a crucial step in preparing the data for modeling by transforming raw data into meaningful features that capture relevant information about employee attrition. This module involves techniques such as feature scaling to normalize numerical features, one-hot encoding or label encoding to represent categorical variables, and handling outliers or skewed distributions. Feature engineering may also include creating new derived features through mathematical transformations, interaction terms, or domain-specific knowledge to enhance the predictive power of the model. Feature selection techniques such as correlation analysis, recursive feature elimination, or feature importance analysis may be employed to identify the

most informative predictors of attrition.

3.1.3 MODEL DEVELOPMENT

The model development module focuses on building and training the predictive model using the Random Forest algorithm, a powerful ensemble learning technique. Random Forest is chosen for its ability to handle both numerical and categorical features, capture complex relationships in data, and mitigate overfitting. This module involves splitting the data into training and testing sets, training the Random Forest model on the training data using multiple decision trees, and tuning hyperparameters to optimize model performance. Cross-validation techniques such as k-fold cross-validation or stratified cross-validation may be used to assess the model's performance and generalization ability.

3.1.4 MODEL EVALUATION

This module is responsible for evaluating the performance of the trained Random Forest model using appropriate metrics and techniques. Evaluation metrics commonly used for binary classification tasks include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Techniques such as confusion matrix analysis, ROC curve analysis, and calibration plots may be employed to assess the model's discrimination ability, calibration, and overall predictive performance. The model's performance may be compared against baseline models or alternative algorithms to determine its effectiveness in predicting employee attrition.

3.1.5 DEPLOYMENT

The deployment module focuses on deploying the trained Random Forest model into production for real-time use within HR systems or workflows. This

involves integrating the model into existing infrastructure, developing APIs or user interfaces for interacting with the model, and deploying it on cloud or on-premise servers. Continuous monitoring and maintenance of the deployed model are essential to ensure its performance and reliability over time, including monitoring for concept drift, data quality issues, or model degradation.

3.1.6 INTERPRETATION AND INSIGHTS

The interpretation and insights module focuses on interpreting the results of the predictive model and deriving actionable insights for HR decision-making. This involves analyzing feature importance to identify the key factors influencing employee attrition, visualizing model predictions and decision boundaries, and generating reports or dashboards to communicate findings to stakeholders. Insights derived from the model can inform targeted retention strategies, identify high-risk employees, and prioritize interventions to reduce attrition rates effectively. Continuous feedback loops may be established to incorporate new data and insights into the model and refine retention strategies over time.

CHAPTER 4

OUTPUT SCREENSHOTS

In [11]:
```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
```

In [12]:
```
Emp_data = pd.read_csv("Dataset01-Employee_Attrition.csv")
Emp_data.head()
```

Out[12]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Department	sa
0	0.38	0.53	2	157	3	0	1	0	sales	
1	0.80	0.86	5	262	6	0	1	0	sales	mec
2	0.11	0.88	7	272	4	0	1	0	sales	mec
3	0.72	0.87	5	223	5	0	1	0	sales	
4	0.37	0.52	2	159	3	0	1	0	sales	

In [13]:
```
Emp_data.shape
```

Out[13]:
```
(14999, 10)
```

In [14]:
```
Emp_data.columns
```

Out[14]:
```
Index(['satisfaction_level', 'last_evaluation', 'number_project',
       'average_montly_hours', 'time_spend_company', 'Work_accident', 'left',
       'promotion_last_5years', 'Department', 'salary'],
      dtype='object')
```

In [15]:
```
Emp_data.dtypes
```

Out[15]:
```
satisfaction_level       float64
last_evaluation          float64
number_project             int64
average_montly_hours       int64
time_spend_company         int64
Work_accident              int64
left                       int64
promotion_last_5years      int64
Department                object
salary                    object
dtype: object
```

In [16]:
```
Emp_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_montly_hours   14999 non-null  int64
 4   time_spend_company     14999 non-null  int64
 5   Work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
 8   Department             14999 non-null  object
 9   salary                 14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

In [17]:
```
Emp_data[Emp_data.duplicated()]
```

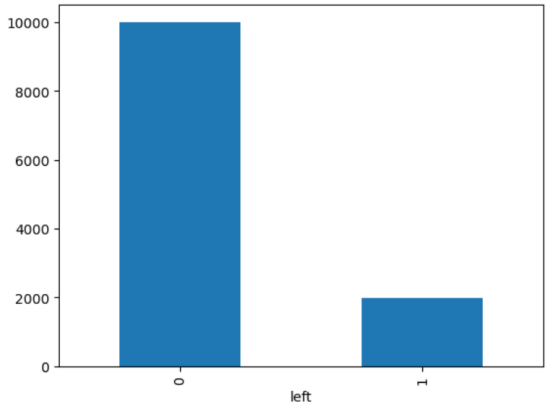
Out[17]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Department
396	0.46	0.57	2	139	3	0	1	0	sales
866	0.41	0.46	2	128	3	0	1	0	accounting
1317	0.37	0.51	2	127	3	0	1	0	sales
1368	0.41	0.52	2	132	3	0	1	0	RandD
1461	0.42	0.53	2	142	3	0	1	0	sales
...	...	...	...	...	...	...	...	...	...
14994	0.40	0.57	2	151	3	0	1	0	support
14995	0.37	0.48	2	160	3	0	1	0	support
14996	0.37	0.53	2	143	3	0	1	0	support
14997	0.11	0.96	6	280	4	0	1	0	support
14998	0.37	0.52	2	158	3	0	1	0	support

3008 rows × 10 columns

```
In [19]: Emp_data1 = Emp_data.drop_duplicates()
         Emp_data1.shape

Out[19]: (11991, 10)

In [20]: Emp_data1.isnull().sum()

Out[20]: satisfaction_level       0
         last_evaluation          0
         number_project           0
         average_montly_hours     0
         time_spend_company       0
         Work_accident            0
         left                     0
         promotion_last_5years    0
         Department               0
         salary                   0
         dtype: int64

In [21]: Emp_data1['left'].value_counts()

Out[21]: left
         0    10000
         1     1991
         Name: count, dtype: int64
```

```
In [22]: Emp_data1['left'].value_counts().plot(kind = 'bar')

Out[22]: <Axes: xlabel='left'>
```

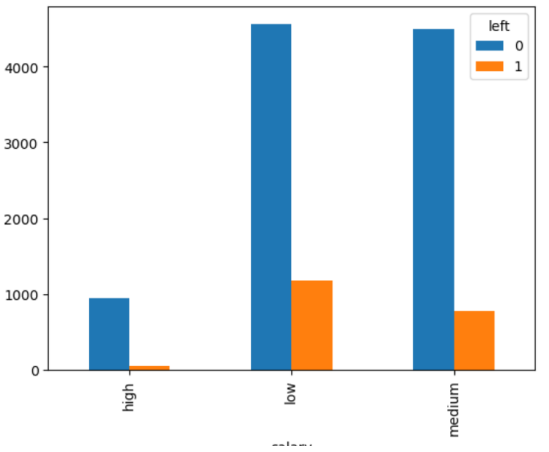



```
In [23]: Emp_data1.head()
```

Out[23]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Department	sa
0	0.38	0.53	2	157	3	0	1	0	sales	
1	0.80	0.86	5	262	6	0	1	0	sales	med
2	0.11	0.88	7	272	4	0	1	0	sales	med
3	0.72	0.87	5	223	5	0	1	0	sales	
4	0.37	0.52	2	159	3	0	1	0	sales	

```
In [25]: pd.crosstab(Emp_data1.salary, Emp_data1.left).plot(kind = 'bar')

Out[25]: <Axes: xlabel='salary'>
```



In [30]:
```
num_feature_list1 = [f for f in Emp_data1.columns if Emp_data1.dtypes[f] == 'float64']
num_feature_list1
```

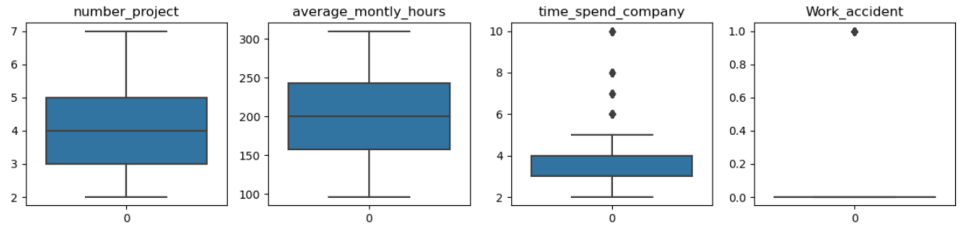
Out[30]:
```
['satisfaction_level', 'last_evaluation']
```

In [31]:
```
num_feature_list2 = [f for f in Emp_data1.columns if Emp_data1.dtypes[f] == 'int64']
num_feature_list2
```

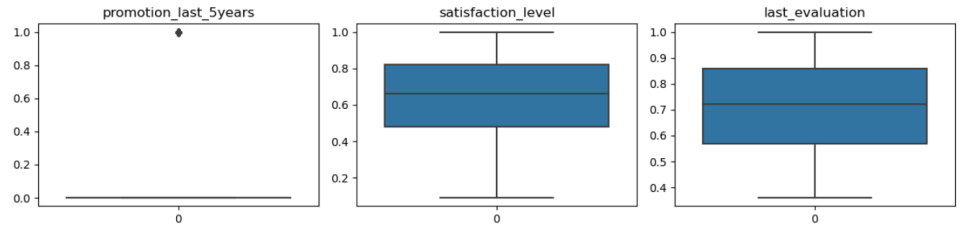
Out[31]:
```
['number_project',
 'average_montly_hours',
 'time_spend_company',
 'Work_accident',
 'left',
 'promotion_last_5years']
```

In [32]:
```
num_col_list = ['number_project',
 'average_montly_hours',
 'time_spend_company',
 'Work_accident',
 'promotion_last_5years','satisfaction_level', 'last_evaluation']
```

In [37]:
```
fig, axes = plt.subplots(ncols = 4, figsize = (12,3))
for column, axis in zip(num_col_list[:4],axes):
    sns.boxplot(data = Emp_data1[column], ax = axis)
    axis.set_title(column)
plt.tight_layout()
plt.show()
```



In [38]:
```
fig, axes = plt.subplots(ncols = 3, figsize = (12,3))
for column, axis in zip(num_col_list[4:],axes):
    sns.boxplot(data = Emp_data1[column], ax = axis)
    axis.set_title(column)
plt.tight_layout()
plt.show()
```



In [112]:
```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
```

In [113]:
```
Emp_data1['salary'] = label_encoder.fit_transform(Emp_data1['salary'])
Emp_data1['Department'] = label_encoder.fit_transform(Emp_data1['Department'])
```

```
C:\Users\amshitha\AppData\Local\Temp\ipykernel_26720\3068568613.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  Emp_data1['salary'] = label_encoder.fit_transform(Emp_data1['salary'])
C:\Users\amshitha\AppData\Local\Temp\ipykernel_26720\3068568613.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  Emp_data1['Department'] = label_encoder.fit_transform(Emp_data1['Department'])
```

In [56]:
```
from sklearn.preprocessing import StandardScaler
std_scaler = StandardScaler()
```

In [68]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

In [69]:
```
x_train.shape
```

Out[69]:
```
(9592, 9)
```

In [70]:
```
x_train.head()
```

Out[70]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	Department	salar
6426	0.86	0.56	5	141	2	0	0	7	
288	0.82	0.91	5	232	5	0	0	9	
5051	0.62	0.52	3	148	3	0	0	7	
11132	0.43	0.74	4	129	6	0	0	3	
3687	0.85	0.79	3	156	3	0	0	7	

In [78]:
```
from sklearn.preprocessing import StandardScaler
std_scaler = StandardScaler()
```

In [121]:
```
xtrain_scaled = std_scaler.fit_transform(x_train)
xtest_scaled = std_scaler.transform(x_test)
```

In [122]:
```
xtrain_scaled
xtest_scaled
```

Out[122]:
```
array([[-2.22616534,  0.4312433 ,  1.89275291, ..., -0.13311211,
         0.39533766, -0.56181526],
       [ 0.4636721 ,  1.44221642,  0.17048512, ..., -0.13311211,
         0.74073148,  1.0287786 ],
       [ 0.7533469 ,  1.50168543,  0.17048512, ..., -0.13311211,
         0.39533766, -0.56181526],
       ...,
       [ 0.7533469 , -0.34185379, -0.69064878, ..., -0.13311211,
         0.39533766,  1.0287786 ],
       [ 0.29814364,  1.56115444,  0.17048512, ..., -0.13311211,
        -0.29544999, -0.56181526],
       [ 0.09123307, -0.10397776, -0.69064878, ..., -0.13311211,
         0.74073148, -0.56181526]])
```

In [123]:
```
from sklearn.ensemble import RandomForestClassifier
Random_forest_model = RandomForestClassifier()
```

In [124]:
```
Random_forest_model.fit(xtrain_scaled, y_train)
```

Out[124]:
```
RandomForestClassifier()
```

In [125]:
```
y_pred = Random_forest_model.predict(xtest_scaled)
y_pred
```

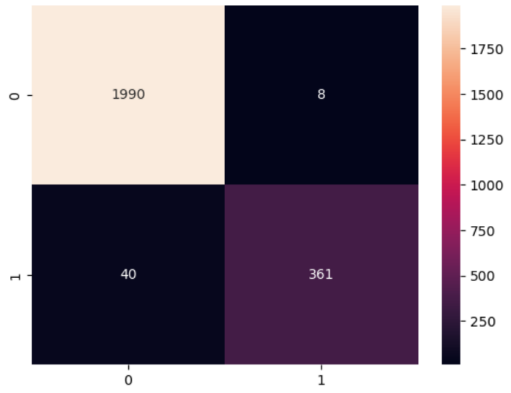
Out[125]:
```
array([1, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [126]:
```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
```

```
[[1990    8]
 [  40  361]]
```

In [127]:
```
sns.heatmap(cm, annot = True, fmt = 'd')
```

Out[127]:
```
<Axes: >
```



In [138]:
```
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
```

In [139]:
```
parameters = {
    'n_estimators' : [50,100],
    'max_features' : ['sqrt', 'log2', None]
}
```

In [140]:
```
grid_search = GridSearchCV(estimator = Random_forest_model,
                           param_grid = parameters)
```

In [143]:
```
grid_search.fit(xtrain_scaled, y_train)
```

Out[143]:

GridSearchCV
estimator: RandomForestClassifier
RandomForestClassifier

In [144]:
```
grid_search.best_params_
```

Out[144]:
```
{'max_features': 'sqrt', 'n_estimators': 100}
```

In [145]:
```
Random_forest_model_new = RandomForestClassifier(max_features = 'sqrt', n_estimators = 50)
```

In [146]:
```
Random_forest_model_new.fit(xtrain_scaled, y_train)
```

Out[146]:

RandomForestClassifier
RandomForestClassifier(n_estimators=50)

In [147]:
```
from sklearn.model_selection import cross_val_score
scores = cross_val_score(Random_forest_model_new, xtrain_scaled, y_train, cv = 5, scoring = 'accuracy')
print('Cross-validation scores =', scores)
```

```
Cross-validation scores = [0.9885357  0.98332465 0.98488008 0.98488008 0.98748697]
```

In [148]:
```
Avg_Model_score = scores.mean()
print('Average Model Score = ', Avg_Model_score)
```

```
Average Model Score =  0.9858214952717489
```

CHAPTER 5

CONCLUSION

In conclusion, the Random Forest algorithm proves to be a powerful tool for predicting employee attrition within an organization. By leveraging a combination of decision trees and ensemble learning, Random Forest can effectively capture complex relationships and interactions within the data, leading to robust predictions. Through the analysis of various features such as employee demographics, performance metrics, and job satisfaction indicators, the Random Forest model can identify patterns indicative of potential attrition risk. By understanding these patterns, organizations can proactively take measures to retain valuable talent, optimize workforce planning, and mitigate the negative impacts of employee turnover. However, it's crucial to acknowledge that predictive models like Random Forest are not foolproof and should be continuously refined and validated with real-world data. Additionally, while predictive analytics can provide valuable insights, they should be complemented with qualitative assessments and strategic HR initiatives to address underlying issues and foster a positive work environment. Overall, the application of Random Forest for employee attrition prediction represents a significant step towards proactive talent management and organizational sustainability.

REFERENCES

[1] Abassi SM, Hollman KW (2000). "Turnover: the real bottom line", Public Personnel Management, 2 (3):333-342.

[2] Bass & Avolio (1993), Transformational leadership and organizational culture. Public Administration Quarterly, 17(1), 112-121.

[3] Denvir, A. & McMahon, F. (1992) “Labour Turnover in London Hotels and the Cost-Effectiveness of Preventative Measure,” International Journal of Hospitality Management, Volume 11, No. 2, pp 143 – 154.

[4] Firth L, David J Mellor, Kathleen A Moore & Claude Loquet (2007). “How can managers reduce employee intention to quit?” J. manage. Psychol. 19 (2): 170-187.

[5] Forbes, A. (1971) “Non-parametric Methods of Estimating the Survivor Function,” The Statistician, vol. 20, pp. 27 – 52.

[6] Goleman (2001), Boyatzis, Richard; McKee, Annie. Primal Leadership: The Hidden Driver of Great Performance, Harvard Business School Press.

[7] Griffeth RW, Hom PW, Gaertner S (2000). "A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium", J. Manage. 26 (3): 463-88.

[8] Herzberg, Frederick (January-February 1968). "One More Time: How Do You Motivate Employees?". Harvard Business Review 46 (1): pp. 53–62.

[9] Hom P.W., Griffeth R.W. (1995). Employee turnover, South-Western college publishing, Cincinnati, OH pp. 200-340.