Project Report Format

Team ID	PNT2022TMID00422
Project Name Project -	Corporate Employee Attrition Analytics
Team Leader	Arun R
Team Members	Akash P,Balaji P,Jeevanandham P

1. INTRODUCTION

1.1 Project Overview

An Data Analytical model is created for the organizations to understand the factors or causes of increasing employee attrition.This model will also provide the areas where the organization could improve for employee retention.

1.2 Purpose

The purpose of this project is to create a data analytical model which can predict the attrition rate for each employee and find out the factors which lead an employee to leave the organization.

It will help organizations to understand the reasons for attrition to decrease attrition or plan in advance the hiring of the new candidate.

It's important to decrease attrition rate because costs associated with losing valuable employees whom you'd like to retain can be staggering.

This scope of the project extends to companies of midsize to large size companies and to all industries.

2. LITERATURE SURVEY

2.1 Existing Problem

Employee attrition is a recurring problem for the HR department. Employee turnover has recently increased dramatically. It is critical for employers to understand whether their employees are unsatisfied or have other reasons for leaving. It is usually prudent to explore the main source of a condition before taking drastic action.
Employees nowadays are more ready than ever to jump from one business to another in searching for a better chance. Employee attrition has become a critical issue in the majority of enterprises.There is no accurate way to determine the root cause of the issue or to deal with it.

2.2 References

1. Predict employee attrition by using predictive analytics - **Ramakrishnan Raman, Sandeep Bhattacharya, Dhanya Pramod.**

2. Employee Attrition Analysis Using Predictive Techniques - **Devesh Kumar Srivastava, Priyanka Nair.**

3. Kishori Singh,Reetu Singh, 2019:A Study on Employee Attrition:Effects and Causes

4. Rama Krishna Garigipati (Koneru Lakshmaiah Education Foundation, India), Kasula Raghu (Mahatma Gandhi Institute of Technology, India) and K. Saikumar (Koneru Lakshmaiah Education Foundation, India),2022:Detection and Identification of Employee Attrition Using a Machine Learning Algorithm

5. Saswat Barpanda and Athira S,2022:Cause of Attrition in an Information Technology-Enabled Services Company: A Triangulation Approach

6. Sabha Yousuf Khan,2019:Study on the Most Determining Factor of Employee Attrition I.E. Age Factor

7. M.S.Kamalaveni, S.Ramesh, T.Vetrivel,2019: A review of literature on employee

retention

2.3 Problem Statement Definition

- Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business,

including the cost of business disruption, hiring new staff and training new staff. As such, there is great business interest in understanding the drivers of, and minimizing staff attrition.

- Obviously, it's important for organization to reduce the main reasons employees Leave for others position. Good people don't leave good organization they leave poor managers.
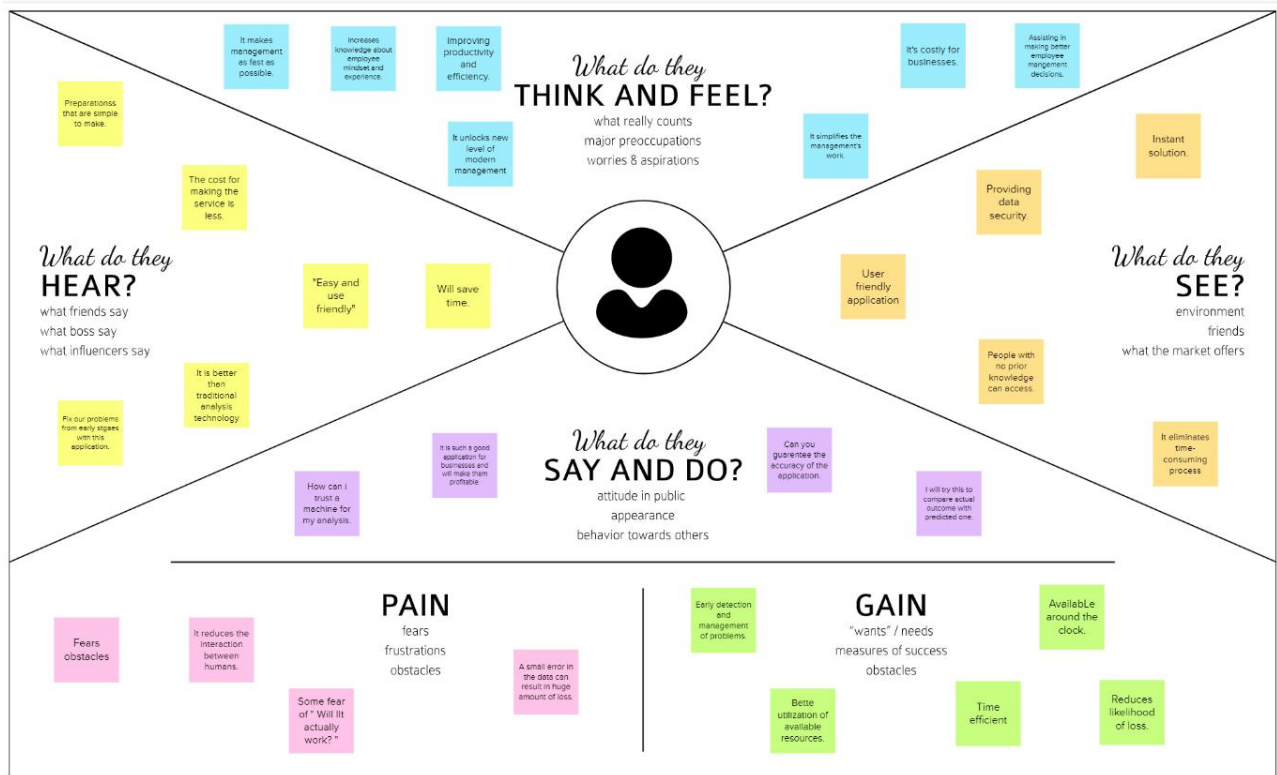- If all employees stay in the same organization for a very long time, most of them will beat the top of their pay scale which will result in excessive manpower

costs.

- When certain employees leave, whose continuation of service would have negatively impacted productivity and profitability of the company, the company is benefited.
- New employees bring new ideas, approaches, abilities & attitudes which can keep the organization from becoming stagnant.
- There are also some people in the organization who have a negative and demoralizing influence on the work culture and team spirit. This, in the long-term, is detrimental to organizational health.
- The first step to building an employee retention model is to determine who is leaving the organization, when they are leaving, and why they are leaving.
- To predict future patterns, we first look to the past to answer the who, when, and why questions. As we noted in a previous article, we can find the answers to these questions by using engagement survey data collected six months to one year in the past, and creating a post-hoc demographic of employees who left the organization voluntarily. Analyzing this demographic will reveal information about turnover in various job roles, tenure levels, business units, and locations – and reveal pockets of high turnover – to tell us who is leaving and when.
- An employee listening perspective will answer the question of why. We can look at what employees who left were telling us about the workplace, work relationships, and their sense of connection to the organization in the months before they left. The comparison of engagement survey data to termination data can reveal areas of the employee experience in need of improvement. We can also look at how the responses of employees who left the organization varied from those who stayed to see which factors in the experience might have been barriers to engagement. This method can be used by any organization that conducts engagement surveys and has the ability to group employees by various demographic factors.
- Exit surveys are another potential data source that can provide richer information. Comparing responses on exit surveys to employees' engagement survey responses can reveal how the employees' perceptions changed over time. Correlating exit and engagement survey data can yield additional capability to predict attrition risk.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming

Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

Define your problem statement

Mandate: **How might we analyze the corporate employee attrition**

Brainstorm

Yogitha Vijayakumar
- Understanding the factors helps to find the right employee attrition rate
- Decide whether to use to supervised or unsupervised learning

Shiva Sankari C
- Evaluating the employee strengths and weaknesses, requirements from the predictive model
- To identify the human resource and the contextual factors that catalyze employee attrition using Triangulation Approach

Sneha Priyadharshini
- Predicting the employee attrition rate using Classification Algorithms so we have only 2 possible values either leave or stay
- Reduce the rates by devising suitable retention strategies

Soundharya G
- Try out different predictive models based on the attrition risk parameters
- Use logistic regression to effectively predict the employee attrition

Top 3 Ideas
- Understanding the factors helps to find the right employee attrition rate
- Try out different predictive models based on the attrition risk parameters
- Predicting the employee attrition rate using Classification Algorithms so we have only 2 possible values either leave or stay

Group Ideas

Prioritize

- Decide whether to use to supervised or unsupervised learning
- Understanding the factors helps to find the right employee attrition rate
- Try out different predictive models based on the attrition risk parameters
- To identify the human resource and the contextual factors that catalyze employee attrition using Triangulation Approach
- Predicting the employee attrition rate using Classification Algorithms so we have only 2 possible values either leave or stay
- Use logistic regression to effectively predict the employee attrition

Feasibility

After you collaborate

3.3 Proposed Solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	To help organizations to find and improve factors that cause the employees to leave an organization thereby increasing attrition rate.
2.	Idea / Solution description	The main idea behind the solution is to use the historic data, work culture of previous organizations and surveys to predict the probability of an employee leaving an organization and factors that are responsible.
3.	Novelty / Uniqueness	Current solutions predict if an employee leaves the organization based only on the employee's historical data, but our solution makes use of employers data too thereby increasing the accuracy of the figured out factors.
4.	Social Impact / Customer Satisfaction	Work force is the major factor contributing to the growth of an organization. Our solution helps the HR team identify the key factors responsible to retain the employees and have increased advantage over their competitors.
5.	Business Model (Revenue Model)	We plan to use the Fee-for-Service model where any organization can use our analytical service for a certain fee. This is a cloud based application.
6.	Scalability of the Solution	This solution can be used by organizations with employees ranging from a few hundreds to thousands. As this is a web based application, computation requirement is covered by SAAS vendor.

3.4 Problem Solution Fit

CUSTOMER SEGMENT(S) 1. HR 2. Talent Acquisition Team 3. Head Hunter Organization	6. CUSTOMER CONSTRAINTS Multitudinous factors that are diffcult to take into consideration for manual analysis.	5. AVAILABLE SOLUTIONS Predict whether an employee will stay in the organization for a period of time.
2. JOBS-TO-BE-DONE / PROBLEMS Develop solution to identify factors resposible for employees to leave an organization	9. PROBLEM ROOT CAUSE 1. Unsatisfactory work life balance 2. Low pay. 3. Toxic working environment 4. No scope for growth.	7. BEHAVIOUR Re-negotiate salary and promotion.
3. TRIGGERS Talented work force leaving the organization to work for the competitors. 4. EMOTIONS: BEFORE / AFTER In Control.	10. SOLUTION Use historic data of employee, previous employer and survey to find the factors resposible.	8. CHANNELS of BEHAVIOUR Offline - Resigning

4. REQUIREMENT ANALYSIS:

Requirement's analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product. These features, called requirements, must be quantifiable, relevant and detailed. In software engineering, such requirements are often called functional specifications and are an important aspect of project management.

4.1 Functional requirements

The aim of our model is to study factors like salary, superior – subordinate relationship, growth opportunities, facilities, policies and procedures, recognition, appreciation, suggestions, co-workers by which it helps to know the Attrition level in the organizations and factors relating to retain them.The models helps to

- To know the satisfactory level of employees towards their job and working conditions.
- To identify the factors which make employees dissatisfied about company's policy and norms.
- To find the areas where companies are lagging behind.

Registration:

- As a HR, I can register for the application by entering my email, password, and confirming my password.
- As a HR, I will receive a confirmation email once I have registered for the application.
- As a HR, I can register for the application through LinkedIn.
- As a HR, I can register for the application through Gmail.
-

Login:

- The HR can log into the application by entering email & password.
- He/She can access the already account from this.

Dashboard:

- The HR can use the dashboard to upload data.
- He/She can update employee data.

Cleaning:

- The HR can clean the data uploaded.
- He/She can avoid the noisy data.

Processing:

The HR can process the input data using a suitable model.

Prediction:

HR can predict the result of employee attrition.

Visualization:

- The HR can use it to visualize the results and see the attrition rate.
- I can see the results in the form of graph,bar etc

4.2 Non-Functional requirements:

Usability:

The proposed system should be easy for the user to operate, enter data, and interpret the output.

Compatibility:

The proposed system should be compatible with all web browsers.

Security:

There is a need for a proper and encrypted login authentication for head chef and admin as employee sensitive information as well as inventory should be protected from hacking.

Flexibility:

If need arises in the future, software can be modified to change the requirements.

Maintainability:

Software can be easily repaired if a fault occurs.

Portability: Software can be easily installed on devices and would run smoothly according to the requirement.

5. PROJECT DESIGN

5.1 Data Flow Diagrams:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.
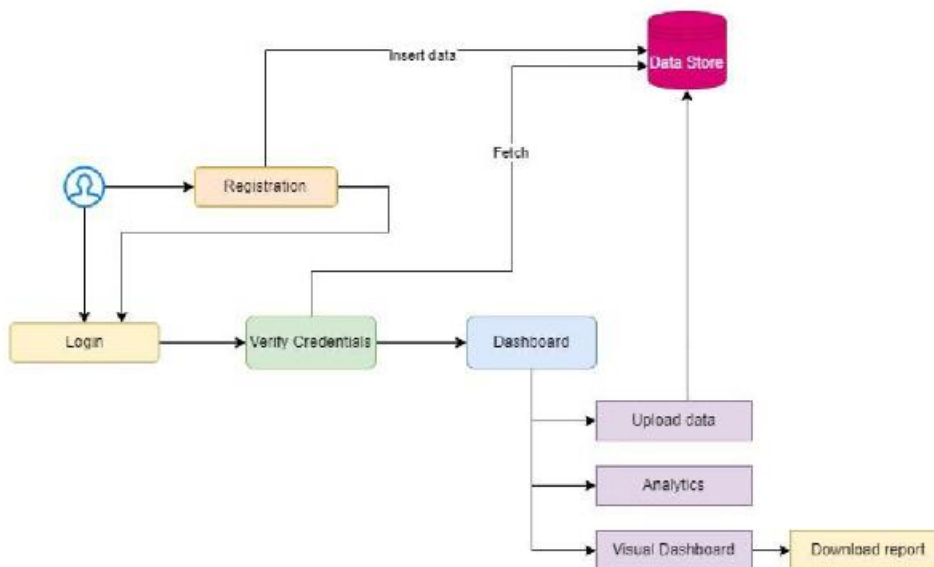
5.2 Solution & Technical Architecture:

Solution architecture is the process of developing solutions based on predefined processes,guidelines and best practices with the objective that the developed solution fits within theenterprise architecture in terms of information architecture, system portfolios, integration requirements and many more.
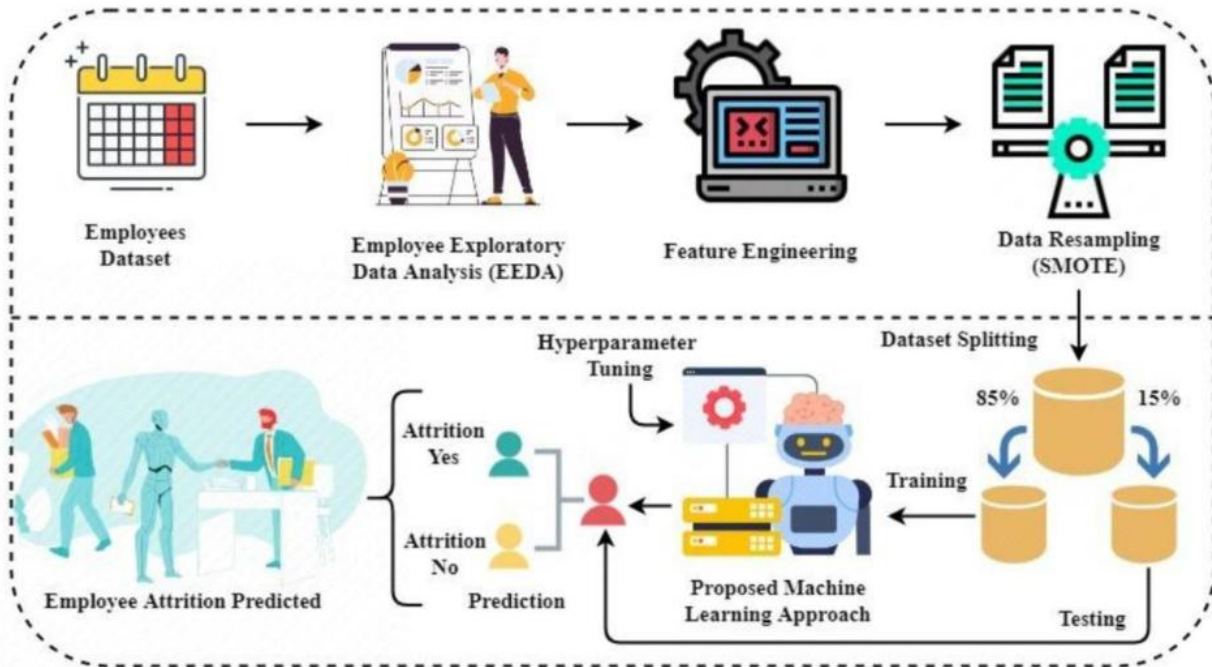
5.3 User Stories:

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Admin	Create Account	USN 1	As an admin, I can create an account.	I can access the account.	High	Sprint 1
	Delete Account	USN 2	As an admin, I can delete an account.	I can delete the account along with its data.	High	Sprint 1
	Modify Credentials	USN 3	As an admin, I can modify an account.	I can update details of the account.	Low	Sprint 1
User	Login	USN 4	Aa a user, i can login to my account.	Access my account	High	Sprint 2
	Upload data	USN 5	I can upload data for analytics.	Upload employee data for analytics.	High	Sprint 2
	Remove data	USN 6	I can remove uploaded data from the server.	Remove uploaded data.	High	Sprint 2
	Modify data	USN 7	I can modify the existing data in the server.	Modify the uploaded data	High	Sprint 2
	View dashboard	USN 8	Perform analytics.	Perform attrition analytics and view the result.	High	Sprint 3
	download report	USN 9	Download the result.	Download the result of analytics.	High	Sprint 4
	Share reports	USN 10	Share the result within the app.	Share the results.	High	Sprint 4
	Manage dashboards	USN 11	Manage multiple dashboards.	Manage multiple analytics.	High	Sprint 4

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	Admin registration	3	High	Arun,Balaji,Akash
Sprint-1		USN-2	As a Admin, I will receive confirmation email on registration.	2	High	Jeevanandham,Balaji, Akash
Sprint-1	Login	USN-3	As a HR, I can log into the application by entering email & password.	5	High	Jeevanandham,Arun, Balaji
Sprint-2	Dashboard	USN-4	As a HR ,I can able access the dashboard	10	High	Akash,Balaji
Sprint-3	Processing	USN-5	As a HR, I can able to process the input data using a suitable model	10	High	Arun,Akash,Balaji
Sprint-4	Predict	USN-6	As a HR, I can able to predict the result of employee attrition	4	High	Arun,Balaji
Sprint- 4	Visualize	USN-7	As a HR, I can able to visualize theresults and see the attrition rate of the employee	6	Medium	Arun,Akash

## 6.2 Sprint Delivery Schedule:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	10	6 Days	25 Oct 2022	30 Oct 2022	10	16 Nov 2022
Sprint-2	10	6 Days	01 Nov 2022	06 Nov 2022	10	18 Nov 2022
Sprint-3	10	6 Days	08 Nov 2022	13 Nov 2022	10	18 Nov 2022
Sprint-4	10	6 Days	14 Nov 2022	19 Nov 2022	10	18 Nov 2022

Velocity:

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

$$Av = Spring\ duration/velocity = 6/10 = 0.6$$

6.3 Reports from JIRA

Burndown Chart
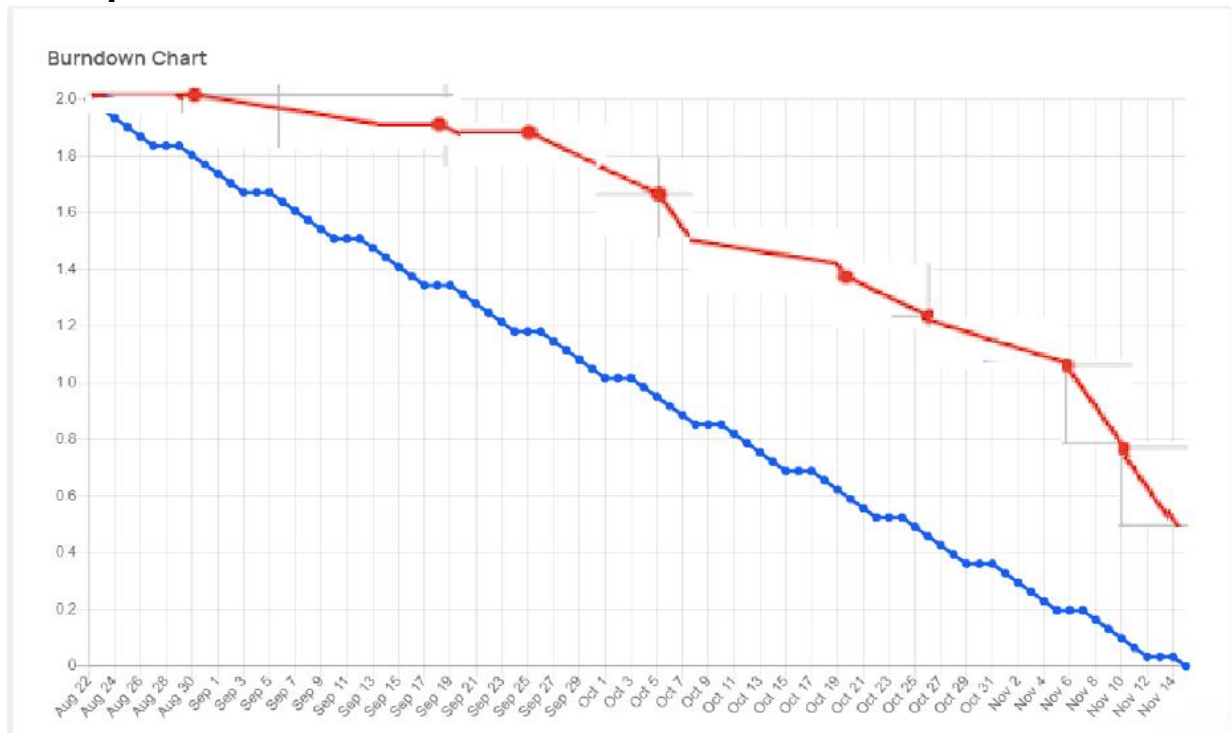
7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

Random Forest Classifier

Random Forest is a powerful algorithm in Machine Learning. It is based on the Ensemble Learning technique (bagging).

```
import pandas as pd
import numpy as np
from lightgbm import LGBMClassifier
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer, make_column_selector
from sklearn.pipeline import make_pipeline
from sklearn. ensemble import RandomForestClassifier, VotingClassifier import joblib
df =
pd.read_csv('general_data.csv')
df.drop(['EmployeeID', 'EmployeeCount', 'Over18','StockOptionLevel'], axis=1, inplace=True) df.drop_duplicates (inplace=True)
X = df.drop('Attrition', inplace=False, axis=1)
y = df.Attrition
ohe = OneHotEncoder (sparse=False, handle_unknown='ignore')
ct = make_column_transformer ((ohe, make_column_selector (dtype_include='object')),
remainder = 'passthrough')
rnd = RandomForestClassifier (n_estimators=500,min_samples_split-80,min_samples_leaf-2, max_features='log2', max_depth=8, random_state=42)
```

7.2 Feature 2

LGBM Classifier

Light GBM is a fast, distributed, high-performance gradient boosting framework that uses a tree-based learning algorithm. It also supports GPU learning and is thus widely used for data science application development.

Light GBM splits the tree leaf-wise with the best fit whereas other boosting algorithms split the tree depth-wise or level-wise rather than leaf-wise. In other words, Light GBM grows trees vertically while other algorithms grow trees horizontally.

```
import pandas as pd
import numpy as np
from lightgbm import LGBMClassifier
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer, make_column_selector from sklearn. pipeline import make_pipeline
from sklearn. ensemble import RandomForestClassifier, VotingClassifier import joblib
df = pd.read_csv('general_data.csv')
df.drop(['EmployeeID', 'EmployeeCount" 'Over18', 'StockOptionLevel'], axis=1, inplace=True)
df.drop_duplicates (inplace=True)
X = df.drop('Attrition', inplace=False, axis=1)
y= df.Attrition
ohe OneHotEncoder (sparse=False, handle_unknown='ignore')
ct = make_column_transformer ((ohe, make_column_selector (dtype_include='object')),
remainder='passthrough')
rnd = RandomForestClassifier (n_estimators=500,min_samples_split-80,min_samples_leaf=2,
max_f features='log2' ,max_depth=8, random_state=42)
1gbm= LGBMClassifier (random_state=42)
vc = VotingClassifier (estimators= [('rnd', rnd), ('1gbm', 1gbm)], voting='soft', n_jobs=-1)
pipe = make_pipeline (ct, vc)
pipe.fit(x, y)
filename = 'model.pkl'
joblib.dump (pipe, filename)
```

8 TESTING

8.1 Test Cases

1) **Login_page**:Verifies the user.

2) **Dashboard:**Verifies the dashboard.

3) **Processing:**Processing of the given data as Intended.

4) **Prediction:**Prediction of data is performed to find the attrtion.

5) **Visualization:**Visualization of data as per requirement.

8.2 User Acceptance Testing
Defect analysis

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	23	3	4	1	31
Fixed	11	2	4	20	37
Not Reproduced	12	12	18	2	44
Skipped	13	23	2	13	51
Won't Fix	12	6	5	4	27
Totals	24	14	13	26	77

Test case analysis:

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	233	56	24	153
Client Application	51	43	1	7
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	12	3	4	5
Final Report Output	5	0	0	5
Version Control	13	0	3	10

9.   ADVANTAGES & DISADVANTAGES

9.1   ADVANTAGES

Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces over fitting problems in decision trees and also reduces the variance and therefore improves the accuracy.

Non linear parameters don't affect the performance of a Random Forest unlike curve based algorithms. So, if there is high non linearity between the independent variables, Random Forest may outperform as compared to other curve based algorithms.

Light GBM uses a histogram-based algorithm i.e it buckets continuous feature values into discrete bins which fasten the training procedure.

9.2   DISADVANTAGES

Random Forest creates a lot of trees (unlike only one tree in case of decision tree) and combines their outputs. By default, it creates 100 trees in the Python sklearn library. To do so, this algorithm requires much more computational power and resources. On the other hand decision tree is simple and does not require so much computational resources
Light GBM splits the tree leaf-wise which can lead to over fitting as it produces many complex trees.

10. CONCLUSION

On the whole, this project was a useful experience. We have gained new knowledge and skills and achieved several of my learning goals. We got insight into professional practice. We learned the different facets of working . We experienced that self exploration, as in many organizations, is an important factor for the progress of projects. Related to our study we learned

more about employee attrition rate prediction and the various approaches and algorithms to achieve the same. There is still a lot to discover and to improve. The methods used at the moment are still not standardized and a consistent method is in development. Furthermore we have experienced that it is of importance of each strategy and how the other one is better than the current algorithm and in which application. We found that the internship is not one-sided, but it is a way of sharing knowledge, ideas and opinions and implementing the same to get results. The internship was also good to find out what our strengths and weaknesses are. This helped me to define what skills and knowledge. We believe that our time spent in learning and surfing regarding various algorithms and the mathematics behind was well worth it and contributed to finding an acceptable solution to build a model and predict the employee's attrition rate. Two main things that we've learned are the importance of time-management skills and self-motivation. At last this project has given us new insights and motivation to pursue a career in the machine learning domain.

11. FUTURE SCOPE

The goal with employee attrition and retention is to strike the right balance of holding on to top talent while accepting that some level of attrition is healthy; employee attrition analytics enables organizations to find that balance.In future this model could take feedback from employees and analyze them to understand what the employees needs to keep the healthy balance intact.New and different models can be used to analyze the data leading to a improved accuracy of the prediction.

Project demo link:
https://drive.google.com/file/d/1mnatPj4eQq0waOYfkCvoc2Oozai0_jDE/view?usp=sharing

Github Repository Link:

https://github.com/IBM-EPBL/IBM-Project-10638-1659194017