Saipriyanka   39110481   Employee promotion prediction

Table of Contents

1.INTRODUCTION

OVERVIEW

The problem of employee promotion has shot to prominence in organizations because of its negative impacts on issues ranging from work place morale and productivity, to disruptions in project continuity and to long term growth strategies.

One way organizations deal with this problem is by predicting the risk of promotion of employees using machine learning techniques thus giving organizations leaders and Human Resources (HR) the foresight to take pro-active action for retention or plan for succession.

However, the machine learning techniques historically used to solve this problem fail to account for the noise in the data in most HR Information Systems (HRIS). Most organizations have not prioritized investments in efficient HRIS solutions that would capture an employee's data during his/her tenure.

One of the major factors is the limited understanding of benefits and cost. It is still difficult to measure the return of investment in HRIS . This leads to noise in the data, which in turn attenuates the generalization capability of these algorithms

2. LITERATURE SURVEY

EXISTING PROBLEM:

Data Retrieval process:
Promotion or career advancement is a process through which an employee of a company is given a higher share of duties, a higher pay scale, or both. Promotion is not just beneficial for employees but is also highly crucial for the employer or business owners. It boosts the morale of promoted employees, increases their productivity, and hence improves upon the overall profits earned by the

organization.

The client is facing a problem in identifying the right people for promotion. The company needs help in identifying the eligible candidates at a particular checkpoint so that they can expedite the entire promotion cycle. This problem can be solved by building a machine learning that automates the process of promoting an employee. we make use of employee datasets to build different classification ML models such as Decision tree, Random forest, KNN, and xgboost. The best model is selected and saved for integration with the flask application.

Proposed Solution:

Model Building:

Now our data is cleaned and it's time to build the model. We can train our data on different algorithms. For this project we are applying four classification algorithms. The best model is saved based on its performance. To evaluate the performance confusion matrix and classification report is used.

DECISION TREE MODEL:

A function named decisionTree is created and train and test data are passed as the parameters. Inside the function, DecisionTreeClassifier algorithm is initialized and

training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.

RANDOM FOREST:

A function named randomForest is created and train and test data are passed as the parameters. Inside the function, RandomForestClassifier algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.

```python
def randomForest(x_train, x_test, y_train, y_test):
    rf = RandomForestClassifier()
    rf.fit(x_train,y_train)
    yPred = rf.predict(x_test)
    print('***RandomForestClassifier***')
    print('Confusion matrix')
    print(confusion_matrix(y_test,yPred))
    print('Classification report')
    print(classification_report(y_test,yPred))
```

KNN MODEL:

A function named KNN is created and train and test data are passed as the

parameters. Inside the function, KNeighborsClassifier algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.

Xgboost Model

A function named xgboost is created and train and test data are passed as the parameters. Inside the function, GradientBoostingClassifier algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.

3.THEORITICAL ANALYSIS:

Block Diagram:
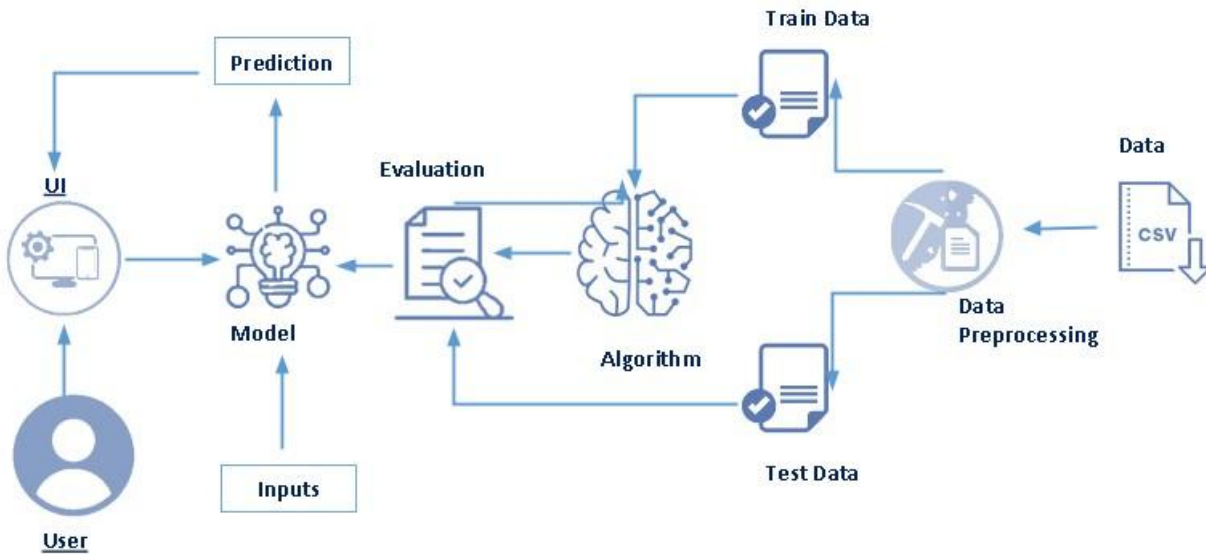
HARDWARE/SOFTWARE DESIGNING:

Requirements are the minimal configuration of a device and software required for the model to work properly and efficiently.

Hardware Requirement:

System : Pentium i3 Processor.
Hard Disk : 500 GB.
Monitor : 15'' LED
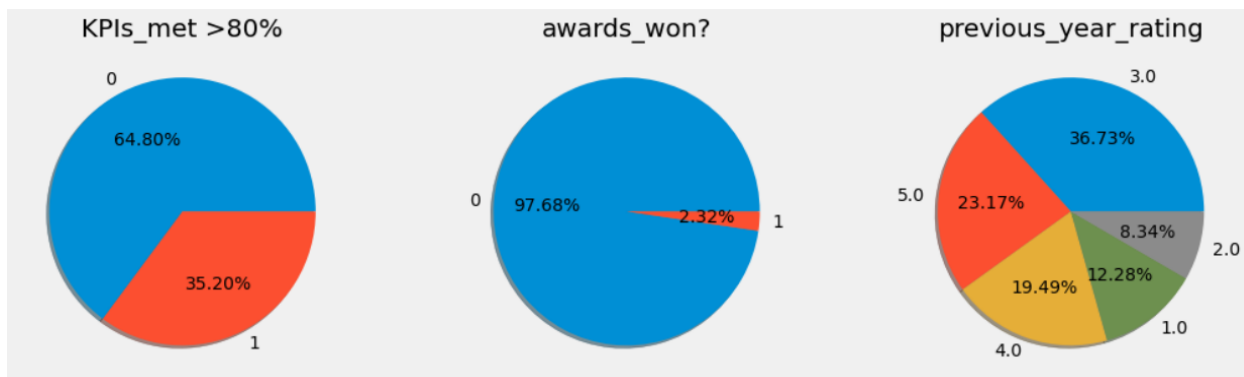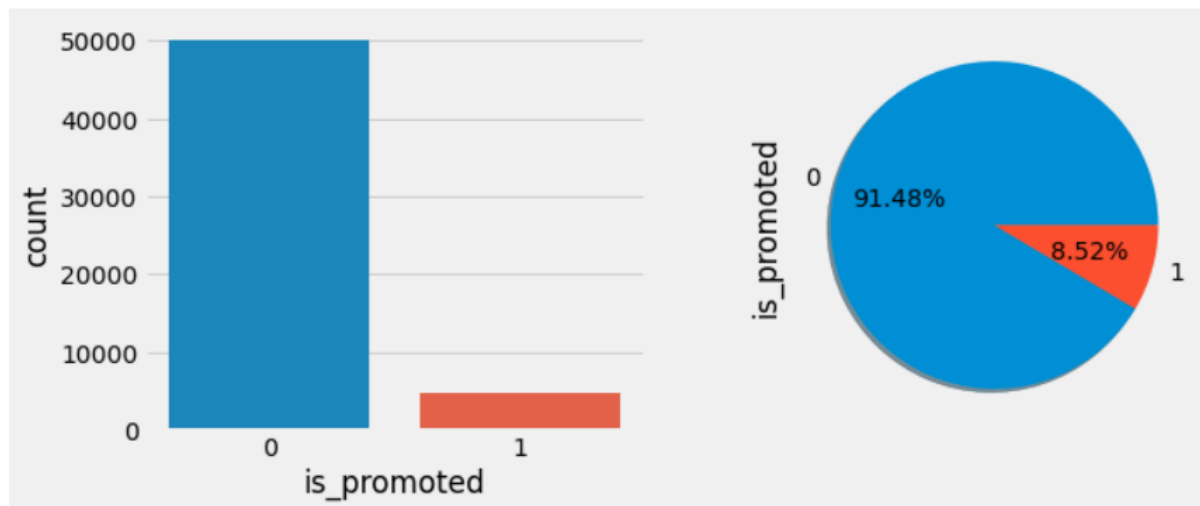Input Devices : Keyboard, Mouse

Ram : 2 GB

Software Requirement:

Operating system : Windows 10.

Coding Language : Python

4.EXPERIMENTAL INVESTIGATIONS

UNIVARIATE ANALYSIS:

In simple words, univariate analysis is understanding the data with single feature.

Here we have displayed two different graphs such as pie plot, box plot and count
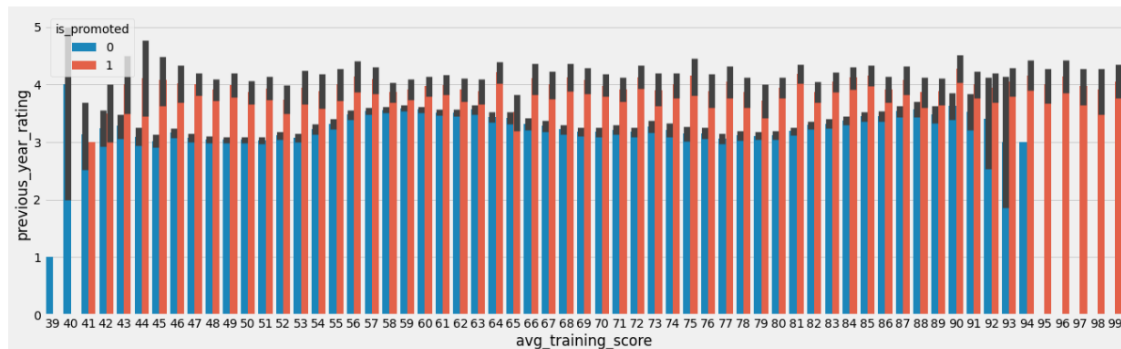
plot.

- Count plot and pie plot are used on the target variable. From the below image, we identified our data is imbalanced. 91% of the employees are not promoted. To get better model performance, imbalanced data should be converted to balanced data. Handling imbalanced data will be discussed on data pre processing.

MULTIVARIATE:

Multivariate analysis

```
In [9]:  """ From the below bar plot, we came to know that employee with training score > 95 & previous year rating > 3 got promoted. """

plt.figure(figsize=(20,6))
sns.barplot(df['avg_training_score'],df['previous_year_rating'],df['is_promoted'])
```

```
Out[9]:  <AxesSubplot:xlabel='avg_training_score', ylabel='previous_year_rating'>
```


In simple words, multivariate analysis is to find the relation between multiple features. Here we have used barplot from seaborn package.

Three features are passed as parameters for barplot(). A clear pattern is understandable from the below plot. Employees with an average training score greater than 95 and a previous year rating greater than 3 got promotions.

Descriptive Analysis

Descriptive analysis is to study the basic features of data with the statistical

process. Here pandas has a worthy function called describe. With this describe

function we can understand the unique, top and frequent values of categorical

features. And we can find mean, std, min, max and percentile values of continuous

features.

```
df.describe(include='all')
```

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	
count	54808.000000	54808	54808	52399	54808	54808	54808.000000	54808.000000	50684.000000	54808.000000	548
unique	NaN	9	34	3	2	3	NaN	NaN	NaN	NaN	
top	NaN	Sales & Marketing	region_2	Bachelor's	m	other	NaN	NaN	NaN	NaN	
freq	NaN	16840	12343	36669	38496	30446	NaN	NaN	NaN	NaN	
mean	39195.830627	NaN	NaN	NaN	NaN	NaN	1.253011	34.803915	3.329256	5.865512	
std	22586.581449	NaN	NaN	NaN	NaN	NaN	0.609264	7.660169	1.259993	4.265094	
min	1.000000	NaN	NaN	NaN	NaN	NaN	1.000000	20.000000	1.000000	1.000000	
25%	19669.750000	NaN	NaN	NaN	NaN	NaN	1.000000	29.000000	3.000000	3.000000	
50%	39225.500000	NaN	NaN	NaN	NaN	NaN	1.000000	33.000000	3.000000	5.000000	
75%	58730.500000	NaN	NaN	NaN	NaN	NaN	1.000000	39.000000	4.000000	7.000000	
max	78298.000000	NaN	NaN	NaN	NaN	NaN	10.000000	60.000000	5.000000	37.000000	

5.FLOW CHART:

Flow chart of the employee promotion prediction
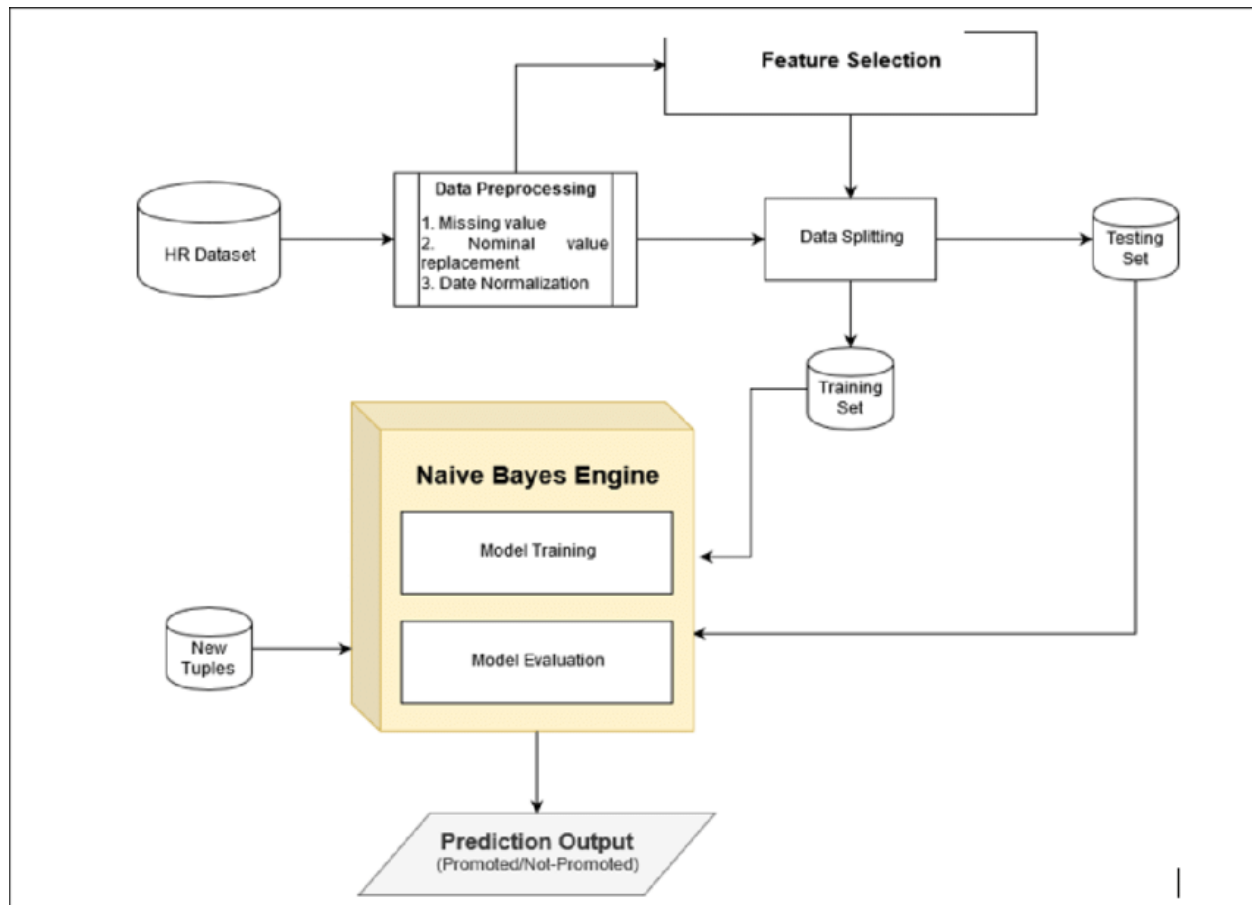
6.RESULT:

For this project,the dataset we will be working with is Employee promotion prediction data set which consists of 54808 employee information and 14 attributes .

The 14 attributes are:

employee_id

departement

region

education

gender

recuirement_channel

no_of_trainings

age

previous _year_rating

length_of_service

kpis_met

awards_won

avg_training_score

is_promoted

7.Advantage

Increasing customer traffic.

The more you promote your brand, the more will the customers know about you and your company and the more will they be interested in your products.

Promotion can be done even by giving out free samples which work wonders for customers

DISADVANTAGE:

No new or fresh ideas are brought into the organization.

The job advertised may require skills not currently available within the organization.

Promotion of an internal employee could cause resentment amongst other employees, who may feel they deserve the post more than the promoted employee.

TESTING MODEL

In this module we test the trained machine learning model using the test dataset Quality assurance is required to ensure that the software system works according

to the requirements. Were all the features implemented as agreed? Does the program behave as expected? All the parameters that you test the program against should be stated in the technical specification document.
Moreover, software testing has the power to point out all the defects and flaws during development. You don't want your clients to encounter bugs after the software is released and come to you waving their fists. Kinds of testing allow us to catch bugs that are visible only during runtime.

increasing customer traffic.
The more you promote your brand, the more will the customers know about you and your company and the more will they be interested in your products. Promotion can be done even by giving out free samples which work wonders for customers

Predictive models are only as good as the data from which they are built, so good data collection practices are crucial to developing high-performing models.
 The data need to be error-free (garbage in, garbage out) and contain relevant information for the task at hand. For example, a loan default model would not benefit from tiger population sizes but could benefit from gas prices over time.
In this module, we collect the employee promotion prediction data from kaggle dataset archives. This dataset contains the information of divorce in previous years.

8.APPLICATIONS:
Employee promotion is is important for employees who work in a company.Generally, based on the performance promotion is given. No organizations will promote their employees by gender, region, and recruitment channel.

9.CONCLUSION:

Our application is able to predict the promotion of a employee.The importance of predicting employee promotion in organizations and the application of machine learning in building models was presented in this paper.

The key challenge of noise in the data from HRIS that compromises the accuracy of these predictive models was also highlighted. Data from the HRIS of a global retailer was used to compare the XGBoost classifier against, six other supervised class that had been historically used to build turnover models.

The results of this research demonstrate that the XGBoost classifier is a superior algorithm in terms of significantly higher accuracy, relatively low runtime and efficient memory utilization for predicting turnover.

its regularization makes it a robust technique capable of handling the noise in the data from HRIS, as compared to the other classifiers, thus overcoming the key challenge in this domain.

Because of these reasons it is recommended to use XGBoost for accurately predicting employee turnover, thus enabling to take actions for retention or succession of employees.

11.REFERENCES:

[1] 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.

[2] M. Stoval and N. Bontis, "Voluntary turnover: Knowledge management – Friend or foe?", Journal of Intellectual Capital, 3(3), 303-322, 2002.

[3] J. L. Cotton and J. M. Tuttle, "Employee turnover: A meta-analysis and review with implications for research", Academy of management Review, 11(1), 55-70, 1986.

[4] L. M. Finkelstein, K. M. Ryanand E.B. King, "What do the young (old) people

imagine me? Content and accuracy of age-based meta stereotypes", European Journal of Work and Organizational Psychology, 22(6), 633-657, 2013.

[5] B. Holtom, T. Mitchell, T. Lee, and M. Eberly, "Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future", Academy of Management Annals, 2: 231-274, 2008