



# **Project On “Employee Promotion Prediction”**

*Submitted By:*

Santosh Kumar Swain  
(1801110099)

Sambit Kumar Behera  
(1801110094)

**Guided By :**

**Dr. Basanta Kumar Swain**  
**Head Of The Department(CSE)**

**DEPT. OF COMPUTER SCIENCE AND ENGINEERING**

**GOVERNMENT COLLEGE OF ENGINEERING  
KALAHANDI, BHAWANIPATNA**

## DECLARATION

We declare that this written submission represents ours ideas in our own words about our topic “**Employee Promotion Prediction**” and where others ideas have been included. We have adequately cited and referenced the original sources. We also declare that we have adhered to all principals of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/act source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has been taken when needed.

Santosh Kumar Swain

Regd. No. 1801110099

Sambit Kumar Behera

Regd. No. 1801110094



**GOVT. COLLEGE OF ENGG. KALAHANDI**  
**BHAWANIPATNA, ODISHA**

---

**CERTIFICATE**

This is to certify that the project entitled “**Employee Promotion Prediction**” submitted by **Santosh Kumar Swain (1801110099)**, **Sambit Kumar Behera (1801110094)** students of 8<sup>th</sup> Semester, Computer Science and Engineering Department, Govt. College of Engineering Kalahandi, Bhawanipatna for the partial fulfilment of the requirement for the award of **Bachelor of Technology in Computer Science and Engineering** degree under **BPUT, Rourkela**, is a record of students’ own study carried out under my supervision and guidance.

This report has not been submitted to any other university or institution for the award of any degree.

Dr. Basanta Kumar Swain  
Head of the Department  
Computer Science and Engg.

## **ACKNOWLEDGEMENT**

The satisfaction that successful completion of this project would be incomplete without the mention of the people who made it possible, without whose constant guidance and encouragement would have made effort go in vain. We consider ourselves privileged to express gratitude and respect towards all those who guided us through the completion of this project.

We convey thanks to our guide Prof. Dr. Basanta Kumar Swain for providing encouragement, constant support and guidance which was of great help to complete this project successfully. .

We would also like to express our gratitude to Dr. Dulu Patnaik Principal, Government College of Engineering Kalahandi, Bhawanipatna for providing us a congenial environment to work in.

Santosh Kumar Swain

Regd. No. 1801110099

Sambit Kumar Behera

Regd. No. 1801110094

## **ABSTRACT**

Promotion is the focus of human resource management research. Because there are few researches about the mining of promotion features in existing studies, this paper uses the data of a Chinese state-owned enterprise, constructs a number of features and applies machine learning methods to predict employee promotion. Firstly, we build personal basic features and post features based on five strategies. Secondly, the correlation analysis is conducted to preliminarily explore the associations between some features and promotion. Then, the model learning and testing are carried out. Experimental results show that the random forest model performs best, which verifies the validity of features. Finally, we calculate the Gini importance of each feature to further analyze its influence on staff promotion. It is found that post features have a higher impact on promotion compared with personal basic features. Among all the features, the working years, the number of different positions and the highest department level greatly affect employee promotion.

## List of Figures

DESCRIPTION	PAGE NUMBER
Fig-1.1 Database	4
Fig-1.2 List of Columns in Train dataset	4
Fig-2.1 Mathematical Formula	12
Fig-2.2 Mathematical Formula	13
Fig-2.3 Mathematical Formula	13
Fig-3.1 Accuracy of AdaBoost	17
Fig-3.2 Accuracy of Gradient Boosting	17
Fig-3.3 Accuracy of Random Forest	17
Fig-3.4 Accuracy of CatBoost	18
Fig-3.5 Accuracy of Light GBM	18
Fig-3.6 Accuracy of XGBoost	18
Fig-3.7 Comparison Graph	19
Fig-3.8 Time Taken Graph	19
Fig-3.9 Columns in Dataset	20
Fig-3.10 Employees are worked across various department	20
Fig-3.11 Percentage of employee got promoted from each department	20
Fig-3.12 Percentage of employee who got promoted from various region	21
Fig-3.13 Distribution of promotion among people	21
Fig-3.14 Percentage of employee who got promotion through RC	22
Fig-3.15 Percentage of employee who got promotion through KPI_met	22
Fig-3.16 Gender wise promotion of Operation department	23
Fig-3.17 Gender wise promotion of Sales department	24
Fig-3.18 Details of employees previous_year_rating	24
Fig-3.19 Distribution of average training score	25
Fig-3.20 Promotion ratio increase with score	25
Fig-3.21 Promotion ratio with respect to age	26
Fig-3.22 Average score of employees from testing data	27
Fig-3.23 Employee promotion with respect to age	27
Fig-3.24 Details of employee who got promotion	28



## TABLE OF CONTENTS

DESCRIPTION	PAGE NUMBER
DECLARATION	i
CERTIFICATE	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
Chapter 1	1-7
1.1 Introduction	1-2
1.2 Related Work	3
1.3 Database	3-4
1.4 Requirement Analysis and Feasibility Study	5
1.4.1 Literature Review	5
1.4.2 Requirement Analysis	6
1.4.3 Functional Requirements	6
1.4.4 Non-Functional Requirements	6-7
Chapter 2	7-17
2.1 Method and Material	7
2.2 Tools and Techonology	7
2.2.1 Python	7
2.2.2 Pandas	8
2.2.3 NumPy	8
2.2.4 Matplotlib	9
2.3 Algorithms	9
2.3.1 AdaBoost	9-10
2.3.2 XGBoost	10-13
2.3.3 Random Forest Classifier	13-14
2.3.4 Gradient Boosting	14-15
2.3.5 Cat Boost	15-16
2.3.6 Light BGM	16-17
Chapter 3	17-28
3.1 Comparison	17-19
3.2 Result	20-28
Chapter 4	29
4.1 Conclusion	29
References	30



# **Chapter 1**

## **1.1 Introduction**

Promotion is not a new term to understand. In today's corporate society everybody wants a salary hike or in other words wants a promotion. Promotion is defined as the Formula applied by the company to get more beneficial outputs by the employee. Promotions play an important role for organizations and individuals. For organizations, it is a way to keep employees committed and motivated towards the company goals by rewarding promoted workers with financial and status gains.

Promotion is the ultimate goal for which an employee works very hard and even do overtime to complete the goals and receive promotion. For individuals, rising through the ranks leads to a boost in morale, wellbeing, and life satisfaction. However, promotions can be a mixed blessing for many – while they provide an increase in occupational status, financial reward, job autonomy, privilege and flexibility, they can often also be accompanied by added responsibility, longer working hours, stress and reduced work-life balance. How it is decided that who will get the promotion?

Previously traditional methods like good attendance, sick leaves, good coordination with co-workers, Benefits received by the company because of that employee, effective education, Background to which that employee belongs are some of the many factors which are taken into consideration before giving promotion to employees and during old times these records are kept in handwritten or printed format which were very difficult to maintain. Then there comes certain problems where if one of the document goes missing then the employee who shall get the promotion will not receive it.

But, today as technology is growing rapidly computers are used almost everywhere including various companies. With the use of computer this process can get lot easier as compared to old times. So now we are going to see a new hassle free way to give promotions to employees who are really entitled for the promotion. Instead of using handwritten or printed documents we will digital documents which are mainly created in Microsoft office, WPS office etc. to store the required information.

Using these technologies we will make a database of the employees which will contain all the required information for giving promotion to all the employees who really deserves it. Then we will use various algorithms to see which algorithm will provide more accuracy then we will use that algorithm to predict the list of the employees who are more likely to get promotion.

Promotions are therefore, a win-some, lose-some game. While workers may win through the status gain, financial and personal growth, they may impact their psychological wellbeing and work-life balance.

## 1.2 Related Works

The majority of prior research related to workplace promotion is focused on an individual's likelihood for being promoted. This topic is studied in relation to both job performance and other related factors. For example, researchers find related factors, such as, personal characteristics, psychological attributes and education level are more related to being promoted than simply performance on the job. Researchers use personality, characteristics, job attributes and psychological information to train machine learning models to predict whether an employee is likely to be promoted.

Use demographic (e.g., gender, date of birth, etc.) and job features (position, department, position type etc.) to predict if an employee is promoted or not using a XGBOOST model reporting an AUC of 0.94. Other research reports a correlation between work related interactions and online social connections with employee promotion. Work related interaction is strongly predictive and correlates with promotion. The authors collect data from an internal social network platform used by the company and train a logistic regression model to predict promotion and resignation of employees.

## 1.3 Database

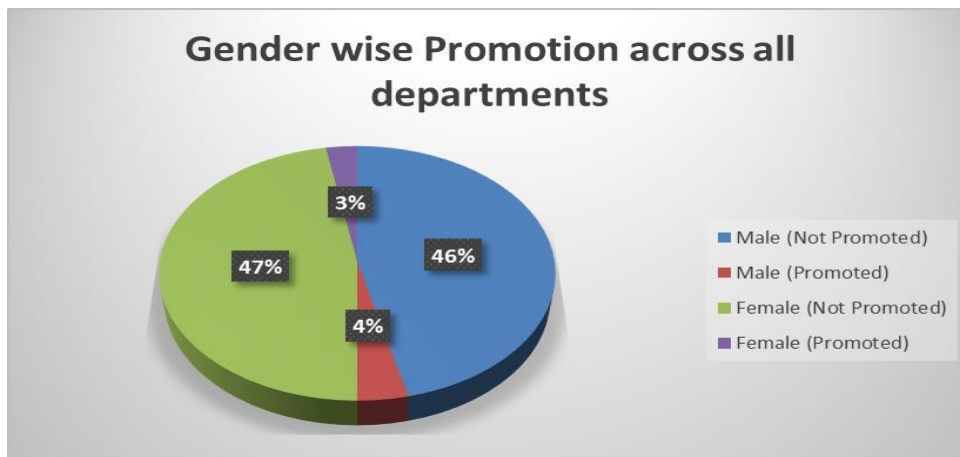
Here we will use two databases named Train and other is Test. Train Dataset is used to see the accuracy of every algorithm used and Test dataset is used to get the final prediction using the algorithm with highest accuracy (XGBoost).

These are certain attributes which are present in the two databases and will be used to make certain graphs and comparing purposes employee\_id, department, region, education, gender, recruitment\_channel, no\_of\_trainings, age, previous\_year\_rating, length\_of\_service, KPIs\_met >80%, awards won, avg\_training\_score, is promoted . Attributes in these two databases are same but the values are different.

Based on the data it is seen that across all the departments if gender is to be compared then,

Male- 0.075771

Female- 0.055802



**Fig-1.1 Database**

As, only 0.05 % of females are to be getting promoted 0.07% males are getting promoted. There are various attributes in the databases but the identifying attribute (uniqueid) is Employees id column.

These database are downloaded from:-

<https://www.mediafire.com/file/vmgcoosabi0jnf8/test.csv/file>

<https://www.mediafire.com/file/s9crcix9plcfcjl/train.csv/file>

Anyone can use this Database as it has been made available to download for all.

Input and Output of using this database

```
df = pd.read_csv('train.csv')
df.columns

[11]
... Index(['employee_id', 'department', 'region', 'education', 'gender',
        'recruitment_channel', 'no_of_trainings', 'age', 'previous_year_rating',
        'length_of_service', 'KPIs_met >80%', 'awards_won?',
        'avg_training_score', 'is_promoted'],
        dtype='object')
```

**Fig-1.2 List of Columns in Train dataset**

## **1.4 Requirement Analysis and Feasibility Study**

This section of the thesis describes the requirements necessary for the project and its feasibility.

### **1.4.1 Literature Review**

Promotion can be used as an incentive tool. It is a way of rewarding the employees for meeting the organizational goals thus it serves as a mean of synchronizing organizational goals with personal goals. Promotion has its importance due to the fact that it carries with it a significant change in the wage package of an employee. Thus, a raise in salary indicates the value of promotion. Promotion follows a defined set pattern which is outlined in the employment bond. In this highly competitive corporate world, promotion can help the competing firms to trace the most productive participant of one organization to be worth hiring for another organization.

In such a way the promotion highlights an employee in the external environment and realizes his worth in the internal environment. According to Carmichael (1983) promotion enhances the yield of an organization when an employee climbs a promotion ladder on the basis of his seniority and resultantly he gets an increased wage rate. However, promotion does not consider to be an incentive device, thus the optimal results cannot be generated by promoting the employee in the organization.

The impact of wage raise, a result of promotion, is found to be more significant than fixed income on job satisfaction. Apart from job satisfaction, the employee satisfaction is determined by satisfaction with promotion. When employees perceive that there are golden chances for promotion they feel satisfied for the respective place in the organization.

### **1.4.2 Requirement Analysis**

In this section, the functionalities need to run the system are described.

### **1.4.3 Functional Requirements**

The system has different functionalities for an User. Their functionalities are described below.

User has the highest privileges among all and is responsible to design the system. They are responsible to take the both training and testing data of the employee. User can view and update the details of employee and compare the dataset with different algorithm. Then user has take one algorithm which has the best fit for predict Promotion.

### **1.4.4 Non-Functional Requirements**

Non-Functional Requirements are the characteristics or attributes of the system that are necessary for the smooth operation of the system.

Those requirements are listed below.

- The system should perform the process accurately and precisely to avoid problems.
- The system should be easy to modify for any updates. Any errors or bugs that are identified should be easy to mend.
- The system should be secure and maintain the privacy of the employees.
- The system should be easy to understand and use.
- Execution of the operation should be fast.

## **Chapter 2**

### **2.1 Method and Materials**

This is the most important section of the thesis. This section describes the detailed workflow of the project and the necessary theoretical background.

### **2.2 Tools and Technologies**

Tools and techniques used in the project are described in this section of the thesis. This project focused was mainly focused on Python Programming and its libraries.

#### **2.2.1 Python**

Python is a high-level object-oriented programming language. It was created by Guido van Rossum in 1991 as Python 0.9.0. It was created as the successor of the ABC programming language. Python 2.0 was released on 16 October 2000 and added many features like list comprehension and garbage collectingsystem. On 3 December 2008, Python 3.0 was released. Python is a very popular programming language and can be used for various purposes. It is widely used for web development, software development, mathematics and data analysis, system scripting, etc. Python is a multi-purpose programming language that works on different platforms like Windows, Linux, Mac, Raspberry Pie, etc. Python is popular than other programming languages because it has a simple syntax than other programming languages. Its syntax allows the programs to write code that is easier to understand and in fewer lines. It runs in an interpreter system. Hence, the code can be executed as soon as it is written.

### **2.2.2 Pandas**

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

### **2.2.3 NumPy**

NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.

NumPy stands for Numerical Python. NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices.

In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy. Arrays are very frequently used in data science, where speed and resources are very important.



### **2.2.4 Matplotlib**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

## **2.3 Algorithms**

An algorithm is simply a set of steps used to complete a specific task. They're the building blocks for programming, and they allow things like computers, smartphones, and websites to function and make decisions. Here we have used 5 different algorithms and they are as follows:-

### **2.3.1 Ada Boost**

Adaptive Boosting which is commonly known as Ada-Boost is a very popular boosting technique. It is an ensemble learning method (also known as “meta-learning”) which was initially created to increase the efficiency of binary classifiers. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers, and turn them into strong ones.

### Advantages:-

1. AdaBoost has a lot of advantages, mainly it is easier to use with less need for tweaking parameters unlike algorithms like SVM. As a bonus, you can also use AdaBoost with SVM.
2. AdaBoost can be used to improve the accuracy of your weak classifiers hence making it flexible.

### Disadvantages:-

1. AdaBoost is also extremely sensitive to Noisy data and outliers so if you do plan to use AdaBoost then it is highly recommended to eliminate them.
2. AdaBoost has also been proven to be slower than XGBoost.

Applications: - It is used to boost the performance of any machine learning algorithm.

### **2.3.2 XGBoost:-**

(Here we will use XGBoost to predict the results as it has given 94.57% accuracy which is best among 5 other algorithms.) It is known as an advanced implementation of gradient boosting algorithm along with some regularized factors. It is also known as the Extreme gradient boosting algorithm and is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular datasets. It is known as an open-source software. Earlier XGBoost is made only for python and R packages but now

it has extended to Java, Scala, Julia and other languages as well. It is known as decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework to do works. In prediction problems involving unstructured data (images, text, etc.) this algorithm's artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

XGBoost algorithm is known to be developed as a research project at the University of Washington. Tianqi Chen and Carlos Guestrin presented their paper at SIGKDD Conference in 2016 and caught the Machine Learning world by fire. Since its introduction, this algorithm has not only been credited with winning numerous competitions but also for being the driving force under the hood for several cutting-edge industry applications. As a result, there is a strong community of data scientists contributing to the XGBoost open source projects with ~350 contributors and ~3,600 commits on GitHub.

#### Advantages:-

1. Fast to interpret
2. Good model performance
3. Good Execution speed

#### Disadvantages

1. Difficult interpretation
2. Overfitting possible if parameters not tuned properly
3. Harder to tune as there are too many hyper parameters.

Mathematical formula:-

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Real value (label) known from the training data-set

Can be seen as  $f(\mathbf{x} + \Delta\mathbf{x})$  where  $\mathbf{x} = \hat{y}_i^{(t-1)}$

**Fig 2.1 Mathematical Formula**

Applications: - Any classification problem. Especially useful if you have too many features and too large datasets, outliers are present, there are many missing values and you don't want to do much feature engineering.

## Methodology

1. We have to find or create the exact database required.
2. Then, we have to install all the modules required through pip install command in python.
3. We will use a dedicated python environment to do this operation 'in this case we are using jupyter notebook'.
4. Then we have to do accuracy test of this program using various algorithms to see which algorithm will give more accuracy.
5. In this case XGBoost has given more accuracy of 94.5% as compared to others whereas other algorithms like Gradient Boosting, Ada Boost, Random Forest has given accuracy results like 94.4%, 93.0%, 93.4%.
6. After getting the best algorithm for this Operation we will use XGBoost to get the results.

Mathematically,

We can represent this model as

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F}$$

**Fig 2.2 Mathematical Formula**

Where, K is the number of trees, f is the functional space of F, F is the set of possible CARTs. The objective function for the above model is given by:

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

**Fig 2.3.2.3 Mathematical Formula**

### **2.3.3 Random Forest Classifier:-**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The reason that the random forest model works so well is: 'A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.'

Advantages:-

1. Random forest can decorrelate trees
2. Reduced error

3. Good Performance on Imbalanced datasets
4. Good handling of missing data
5. Handling of huge amount of data

Disadvantages:-

1. Features need to have some predictive power else they won't work
2. Predictions of the trees need to be uncorrelated.
3. Appears as Black Box

Applications: - Credit card default, fraud customer/not, easy to identify patient's disease or not, recommendation system for ecommerce sites.

### **2.3.4 Gradient Boosting**

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to AdaBoost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is an important parameter used in this technique known as Shrinkage (It refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate which ranges between 0 to 1).

Advantages:-

1. Often provides predictive accuracy that cannot be trumped.
2. Lots of flexibility

3. No data pre-processing required

Disadvantages:-

1. Gradient Boosting Models will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting.
2. Computationally expensive
3. The high flexibility results in many parameters that interact and influence heavily the behavior of the approach (number of iterations, tree depth, regularization parameters, etc.).

Applications: - It is used in regression and classification tasks

### **2.3.5 Cat Boost**

Cat Boost comes from two words, Categorical Boosting. Where Cat signifies Categorical, and Boost means Boosting. It is a famous ensemble machine learning algorithm based on gradient boosting. This algorithm overtakes many other Boosting algorithms like XGBoost, LightGBM, etc., in various aspects like performance, accuracy, implementation, hyper tuning parameters, and many more. Categorical Boosting goes well with categorical data, but it can also handle numerical and text data features.

Advantages:-

1. It gives us great results for categorical data.
2. It can train our model on GPU that significantly increase the speed of learning.

Disadvantages:-

1. It performs only better than other algorithms only when we have categorical data.
2. Can perform very bad if the variables are not properly tuned

Applications: -

1. Weather forecasting
2. Fraud detection
3. Sales forecasting

### **2.3.6 LightGBM**

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.

Advantages: -

1. Faster training speed and higher efficiency
2. Compatibility with Large Datasets

Disadvantages:-

1. Overfitting
2. Complexity with Small Datasets

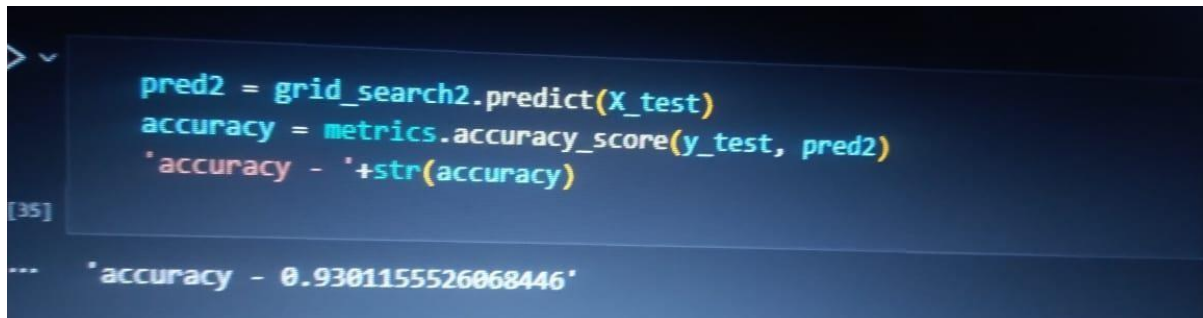
Applications: - used for ranking, classification and many other machine learning tasks.



## Chapter 3

### 3.1 Comparison

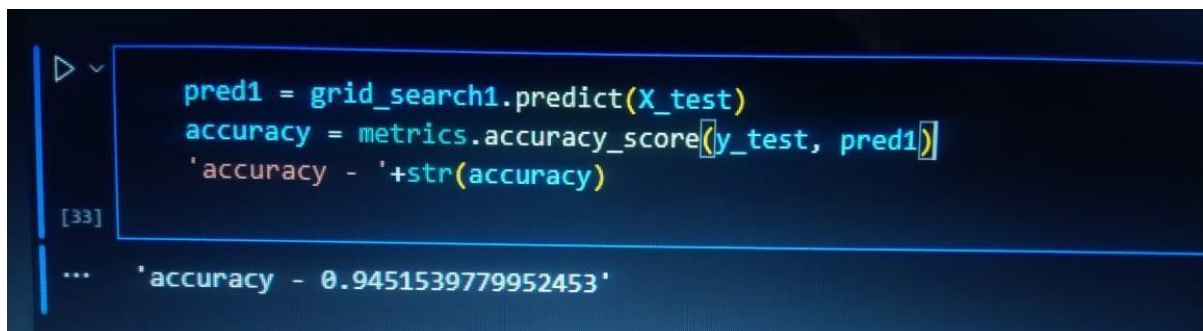
Ada Boost

A screenshot of a Jupyter Notebook cell. The code defines 'pred2' as the prediction from 'grid\_search2' on 'X\_test', calculates 'accuracy' using 'metrics.accuracy\_score(y\_test, pred2)', and prints it. The output shows an accuracy of 0.9301155526068446.

```
> \n\n    pred2 = grid_search2.predict(X_test)\n    accuracy = metrics.accuracy_score(y_test, pred2)\n    'accuracy - '+str(accuracy)\n\n[35]\n\n... 'accuracy - 0.9301155526068446'
```

**Fig-3.1 Accuracy of AdaBoost**

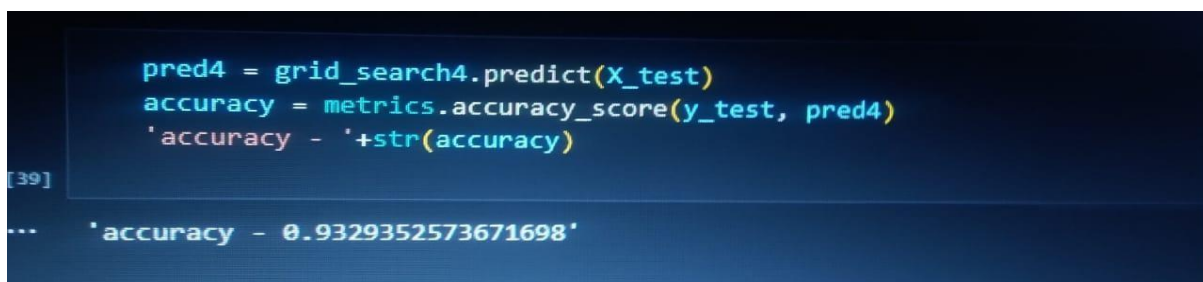
Gradient Boost

A screenshot of a Jupyter Notebook cell. The code defines 'pred1' as the prediction from 'grid\_search1' on 'X\_test', calculates 'accuracy' using 'metrics.accuracy\_score(y\_test, pred1)', and prints it. The output shows an accuracy of 0.9451539779952453.

```
\n\n    pred1 = grid_search1.predict(X_test)\n    accuracy = metrics.accuracy_score(y_test, pred1)\n    'accuracy - '+str(accuracy)\n\n[33]\n\n... 'accuracy - 0.9451539779952453'
```

**Fig-3.2 Accuracy of Gradient Boost**

Random Forest

A screenshot of a Jupyter Notebook cell. The code defines 'pred4' as the prediction from 'grid\_search4' on 'X\_test', calculates 'accuracy' using 'metrics.accuracy\_score(y\_test, pred4)', and prints it. The output shows an accuracy of 0.9329352573671698.

```
\n\n    pred4 = grid_search4.predict(X_test)\n    accuracy = metrics.accuracy_score(y_test, pred4)\n    'accuracy - '+str(accuracy)\n\n[39]\n\n... 'accuracy - 0.9329352573671698'
```

**Fig-3.3 Accuracy of Random Forest**

## Cat Boost

```
pred5 = grid_search5.predict(X_test)
accuracy = metrics.accuracy_score(y_test, pred5)
'accuracy - '+str(accuracy)

[41]
... 'accuracy - 0.945209266323879'
```

**Fig-3.4 Accuracy of Cat Boost**

## Light GBM

```
pred6 = grid_search6.predict(X_test)
accuracy = metrics.accuracy_score(y_test, pred6)
'accuracy - '+str(accuracy)

[55]
... 'accuracy - 0.9446563830375407'
```

**Fig-3.5 Accuracy of Light GBM**

## XGBoost

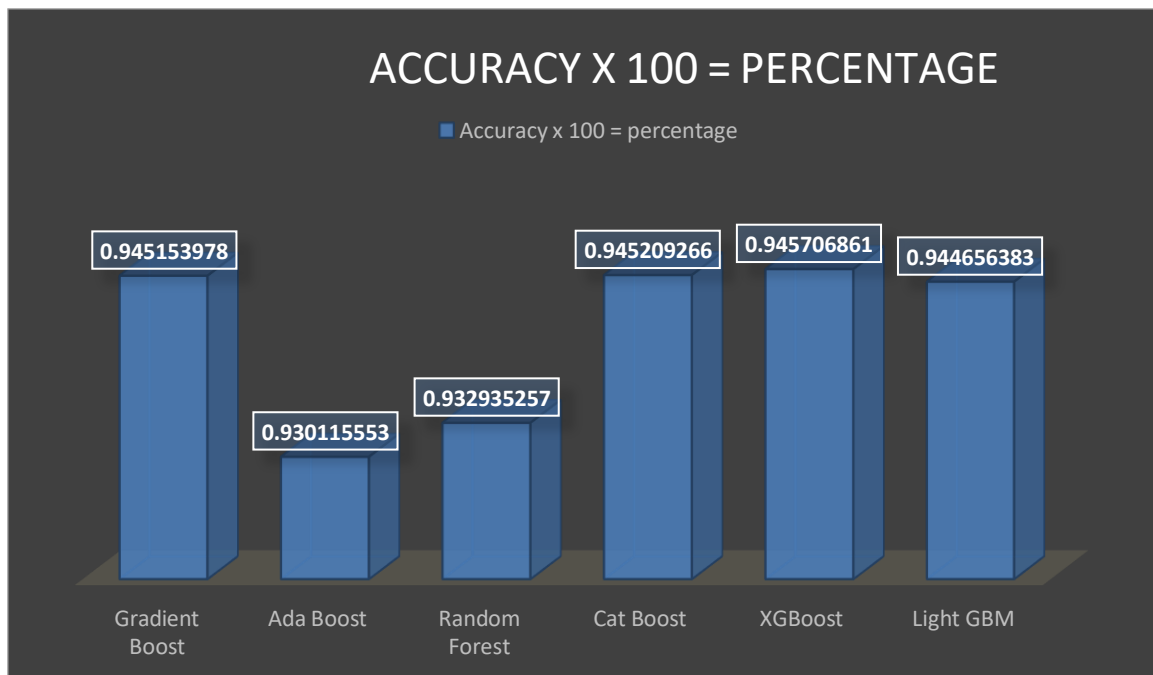
```
pred3 = grid_search3.predict(X_test)
accuracy = metrics.accuracy_score(y_test, pred3)
'accuracy - '+str(accuracy)

[37]
... 'accuracy - 0.9457068612815834'
```

**Fig-3.6 Accuracy of XGBoost**

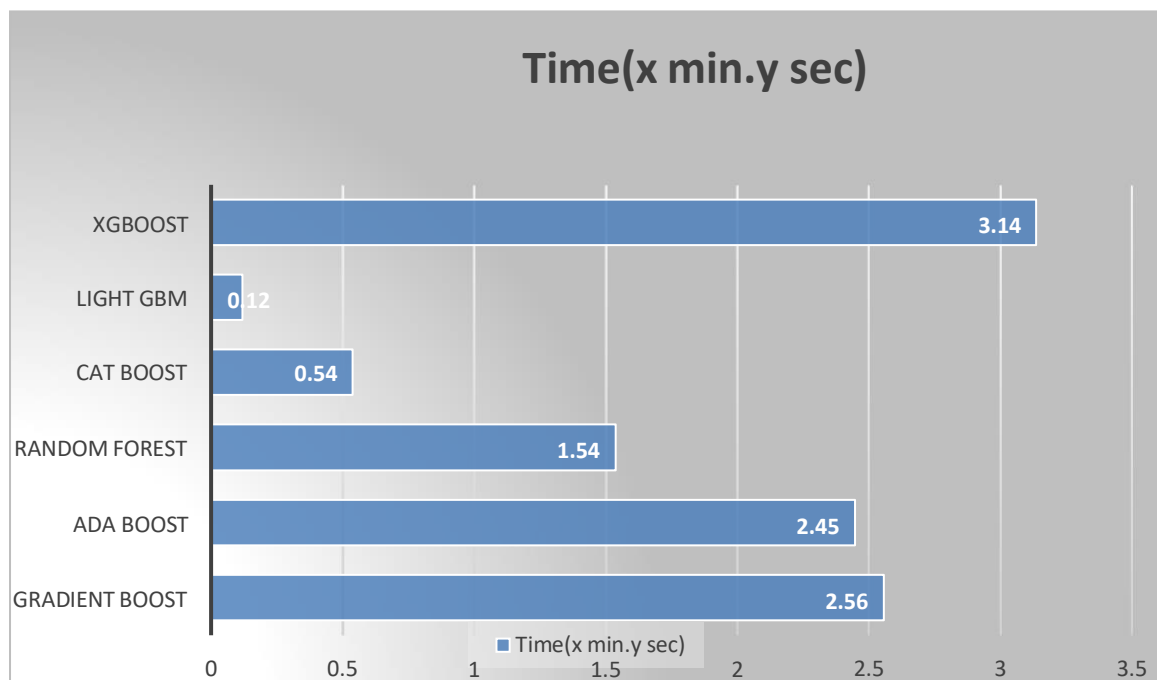
## Comparison graph

Accuracy



**Fig-3.7 Comparison Graph**

Time Taken



**Fig-3.8 Time Taken Graph**

## 3.2 Result

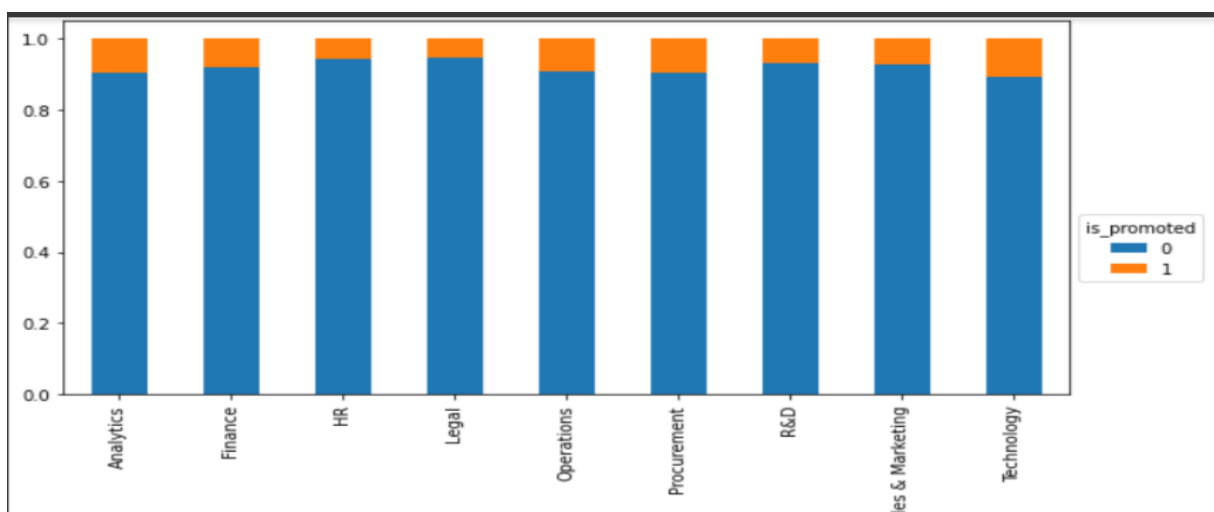
```
Index(['employee_id', 'department', 'region', 'education', 'gender',  
      'recruitment_channel', 'no_of_trainings', 'age', 'previous_year_rating',  
      'length_of_service', 'KPIs_met >80%', 'awards_won?',  
      'avg_training_score', 'is_promoted'],  
      dtype='object')
```

**Fig-3.9 Columns in dataset**

These are the columns that are used for the Employees details.

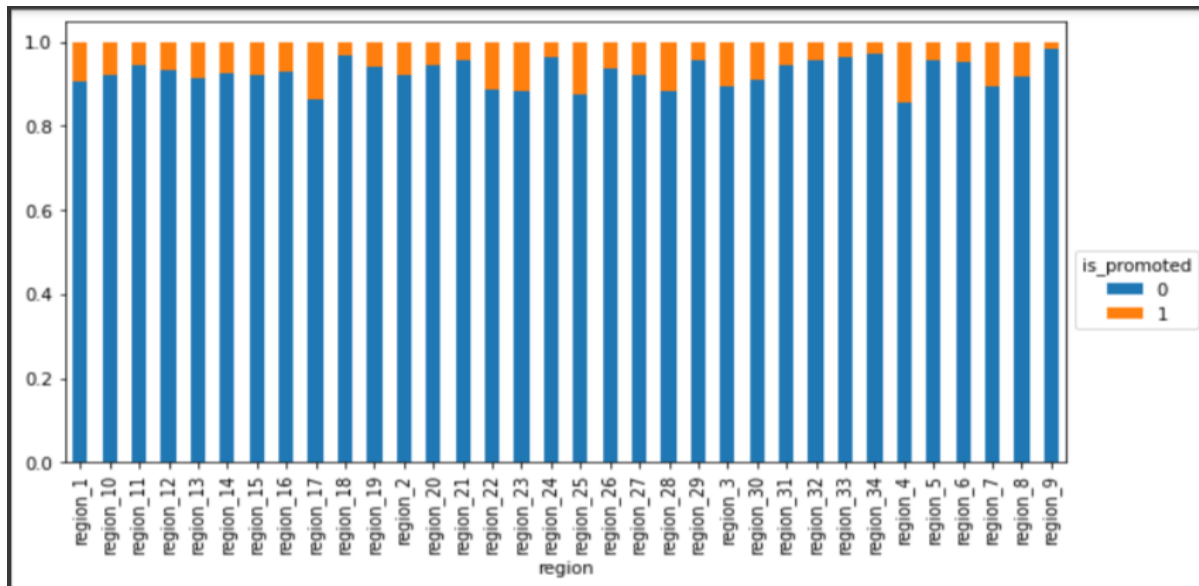
```
↳ Sales & Marketing    16840  
   Operations          11348  
   Technology          7138  
   Procurement         7138  
   Analytics           5352  
   Finance             2536  
   HR                  2418  
   Legal               1039  
   R&D                  999  
   Name: department, dtype: int64
```

**Fig-3.10 Employees are worked across various department**



**Fig-3.11 Percentage of Employee got Promotted From each Department**

Technology department has the highest percentage of employee who got promoted while legal has a least number of percentage who got promoted.



**Fig-3.12 Percentage of employee who got promoted from various region.**

Region\_4 has the highest percentage of promotion while Region\_9 has the least number.

	is_promoted	0	1
education			
Bachelor's		0.917969	0.082031
Below Secondary		0.916770	0.083230
Master's & above		0.901441	0.098559

**Fig-3.13 Distribution of promotion among people with different educational background.**

From Bachelors educational background we have to see that 91.79% of employees are not promoted while 8.2% of employee have got the promotion.

From Below Secondary 91.67% of employees are not promoted but 8.32% of employees have got the promotion.

From Master's & above 90% of employees have not get the promotion but 9.85% of employee who have got promotion.

From this we can see that the percentage are nearly same in different educational background.

is_promoted	0	1
recruitment_channel		
other	0.916048	0.083952
referred	0.879159	0.120841
sourcing	0.914987	0.085013

**Fig-3.14 Employee who got promoted through Recruitment Channel**

From other 91.6% of employees are not promoted but 8.3% of employee have got the promotion.

From reffered 87% of employees are not promoted but 12% of employees have got the promotion.

From sourcing 91.49% have not get the promotion but 8.5% of employees have got the promotion.

As we can see that the employees that are recruited through referral are more likely to be promoted than from other recruitment\_channel.

is_promoted	0	1
KPIs_met >80%		
0	0.960413	0.039587
1	0.830906	0.169094

**Fig-3.15 Percentage of Employee who got promoted through KPI\_met**

From this we have to see that KPIs (Key Performance Indicator) met is greater than 80% is very important for Employees to be promoted.

From KPIs\_met<80% that 96% of employees have not got the promotion, but 3% of employees have got the promotion.

In case on KPIs\_met >80% there are 83% of employee have not got their promotion, but 16% of employee have got the promotion..

As we can see that the employees whose KPIs\_met is greater than 80% have more percentage of promotion that KPIs\_met less than 80%.

is_promoted	0	1
gender		
f	0.905495	0.094505
m	0.912907	0.087093

**Fig-3.16 Genderwise promotion of Operation Department**

From this we have to see that the measurement of percentage of promotion from Operation department with Gender.

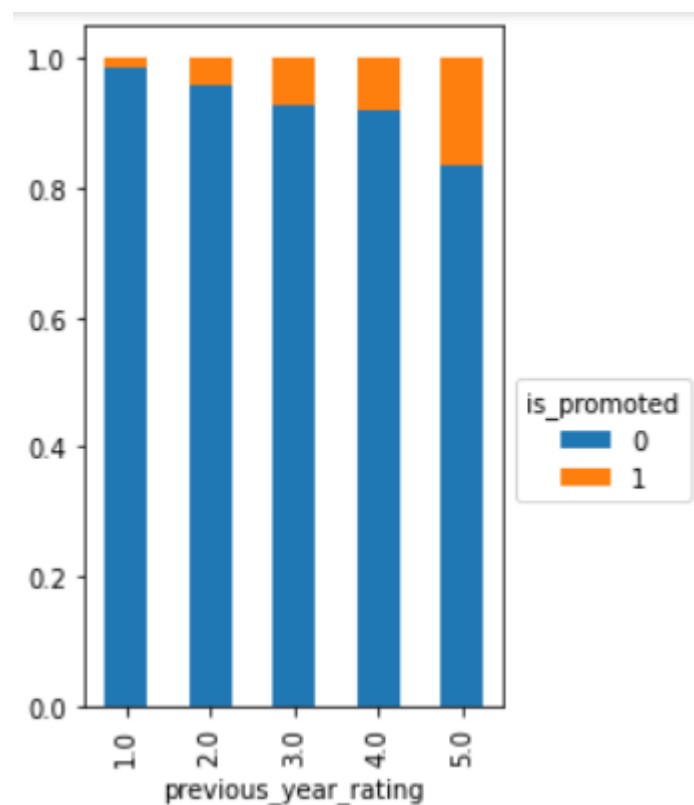
In Operation Department 9.4% of Female employees have got the promotion while 8.7% of Male employees have got the promotion.

As we can see that there is a little bit percentage difference between Male and Female.

is_promoted	0	1
gender		
f	0.944198	0.055802
m	0.924229	0.075771

**Fig-3.17 Genderwise promotion of Sales Department**

This table shows that the comparison between Male and Female from Sales Department that 5% of female have got their promotion while 7% of Male employees have got their promotions.



**Fig-3.18 Details of employees previous\_year\_rating**

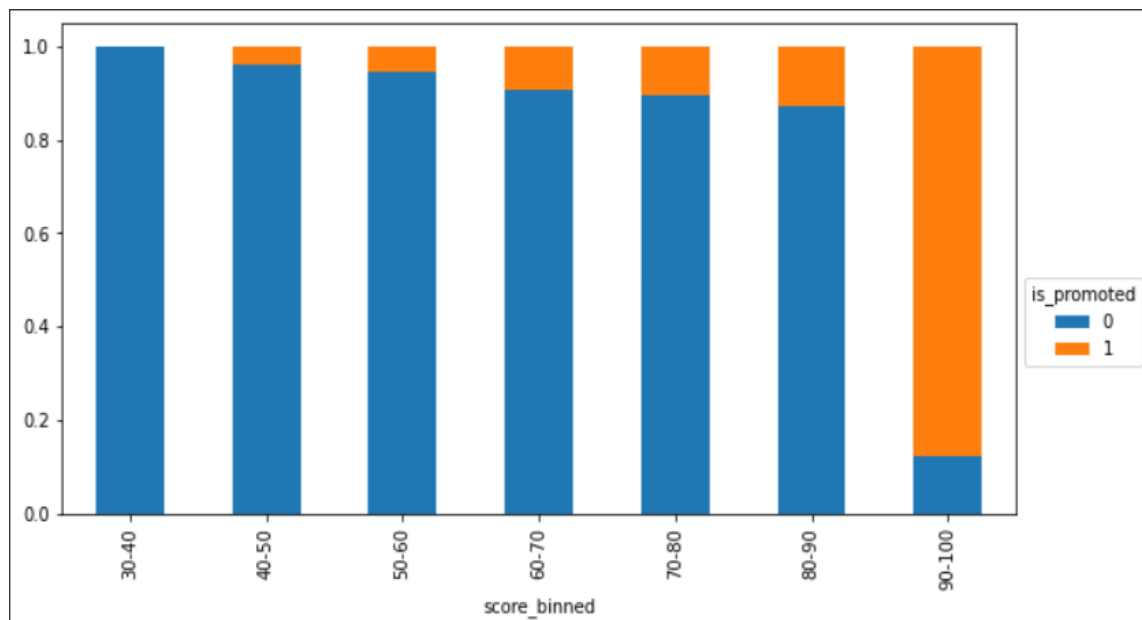
This Graph shows that the details of employees previous\_year\_rating. Whose previous\_year\_rating is 5.0 have more promoted from other rating.



50-60	16020
40-50	11996
60-70	9973
80-90	8739
70-80	7494
90-100	579
30-40	7

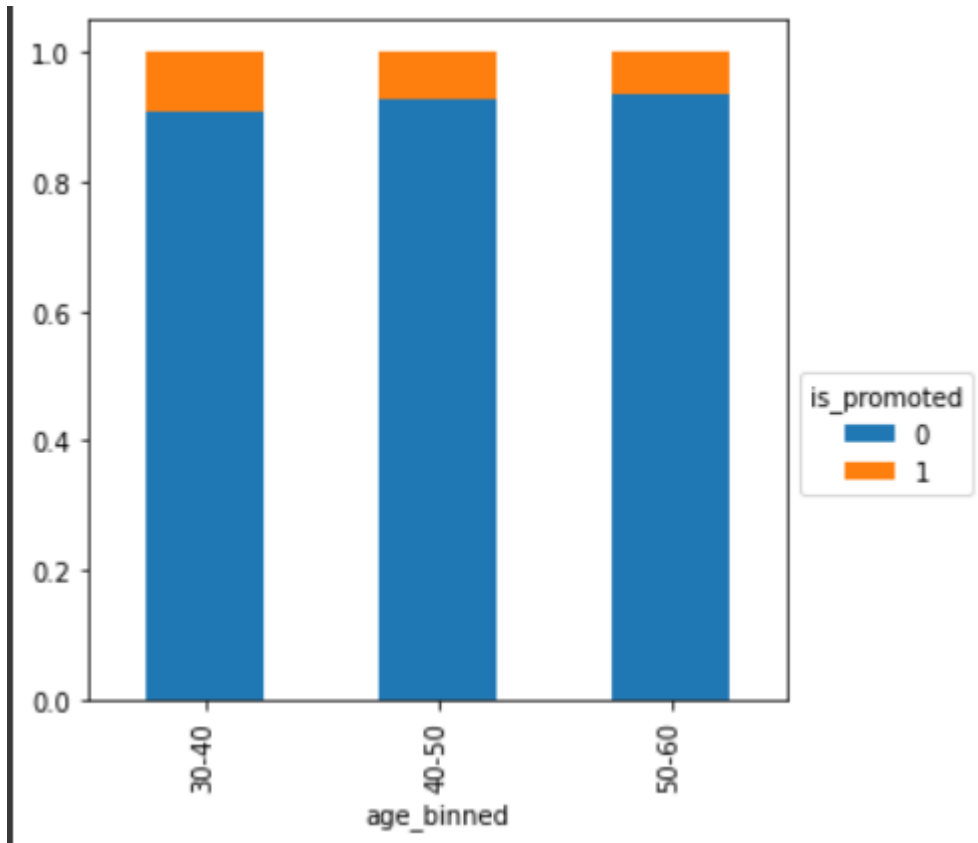
**Fig-3.19 Distribution of average training score**

Most of the employee have score in the range 50-60.



**Fig-3.20 Promotion ratio increases with the score**

In this graph we have to see that Promotion ratio increases with the score and the ratio is very high in the range 90-100 that means getting promoted is highly dependent on the average score.



**Fig-3.21 Promotion ratio with respect to age**

In this graph we have to see that Promotion ratio is high in the age of 30-40. In the age of 50-60 less percentage of employee have got their promotion.

According to our analysis XGBoost algorithm is best for Employee Promotion Prediction.

We are given the dataset of random company employee repository. We have to show which employee have the right credential to get the Promotion. We use XGBoost algorithm to predict the list of employee who are most likely to get Promotion.

The accuracy is high as compared to other algorithm.

```
'accuracy - 0.9457068612815834'
```

50-60	6859
40-50	5104
60-70	4263
80-90	3802
70-80	3214
90-100	242
30-40	6

**Fig-3.22 Average score of Employee From testing data**

The details of the table shows that the average score of employee that have given from Testing Dataset.

30-40	11078
40-50	3355
50-60	1219
60-70	0
70-80	0
80-90	0
90-100	0

**Fig-3.23 Details of Employee promotion with respect to age**

This table shows the details of employee who have the chance to get the promotion with respect to age.

ervice	KPIs_met >80%	awards_won?	avg_training_score	score_binned	age_binned	is_promoted
5	1	0	92	90-100	NaN	1
2	0	0	69	70-80	NaN	1
2	0	0	88	90-100	30-40	1
4	0	1	81	70-80	40-50	1
5	0	0	66	70-80	NaN	1
...	...	...	...	...	...	...
11	0	0	78	40-50	NaN	1

124	55735	Technology	region_31	Master's & above		
139	63901	HR	region_7	Bachelor's	m	other
...	...	...	...	...	...	...
23321	60834	Procurement	region_11	Master's & above	m	sourcing
23382	5452	Sales & Marketing	region_7	Bachelor's	m	other
23387	29019	Analytics	region_22	Bachelor's	m	other
23454	55030	Sales & Marketing	region_28	Master's & above	m	other
23489	5973	Technology	region_17	Master's & above	m	other

690 rows x 16 columns

**Fig-3.24 Employee who have got promotion**

These two table shows that number of employee who have got promotion.

## **Chapter 4**

### **4.1 Conclusion**

Promotion is the ultimate goal for which an employee works very hard and even do overtime to complete the goals and receive promotion. For individuals, rising through the ranks leads to a boost in morale, wellbeing, and life satisfaction. However, promotions can be a mixed blessing for many – while they provide an increase in occupational status, financial reward, job autonomy, privilege and flexibility, they can often also be accompanied by added responsibility, longer working hours, stress and reduced work-life balance. How it is decided that who will get the promotion?

With these technologies we will use various algorithms to see which algorithm will provide more accuracy then we will use that algorithm to predict a list of the employees who are more likely to get promotion. Promotions are a win-some, lose-some game. While workers may win through the status gain, financial and personal growth, they may impact their psychological wellbeing and work-life balance.

## References

1. Sk. Md. Nizamuddin, Advertising and Sales Promotion.
2. Yuxi Long, Jiamin Liu, Ming Fang, Tao Wang, Wei Jiang, "Prediction of Employee Promotion Based on Personal Basic Features and Post Features".
3. Subigya Nepal, Shayan Mirjafari, Gonzalo J. Martinez, Pino Audia, Aaron Striegel, Andrew T. Campbell, "Detecting Job Promotion in Information Workers Using Mobile Sensing"
4. G. S. Thakur, A. Gupta, and S. Gupta, "Data Mining for Prediction of Human Performance Capability in the Software Industry," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 53--64, 2015.
5. Gupta, Sangita & V, Suma. (2014). Empirical Study on Selection of Team Members for Software Projects - Data Mining Approach.
6. Lai, H. H. 2012. Study on influence of employee promotion system on organizational performance. International Journal of organizational Innovation. 5, 1 (July. 2012), 231—251
7. B. Ragins, and E. Sundstrom (1989). Gender and power in organizations: A longitudinal perspective. Psychological Bulletin, vol. 105, pp.51—88
8. Jason Brownlee, XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn.
9. Magnus Lie Hetland, Python Algorithms: Mastering Basic Algorithms in the Python Language
10. Jose M Cortina and Joseph N Luchman. 2012. Personnel selection and employee performance. Handbook of Psychology, Second Edition 12 (2012).