**Report**

**On**

**Prediction of Employee Success using Data Science**

**By**

**VANSH JAIN**

**IDNO: 2020A7PS0079U**

**Under the supervision of**

**Dr. Siddhaling Urolagin**

**Lecturer**

**Submitted in fulfilment of the requirements of**

**CS F376 DESIGN PROJECT**

****

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**DUBAI CAMPUS, DUBAI UAE**

**FEBRUARY 2023 – MAY 2023**

**ACKNOWLEDGEMENTS**

I would like to express my deepest sense of gratitude, first and foremost, to my Supervisor **Dr. Siddhaling Urolagin**, Lecturer, Computer Science Department, BITS Pilani, Dubai Campus, United Arab Emirates, for his valuable guidance and encouragement during this project. I am extremely grateful for his able guidance, valuable technical inputs, and useful suggestions.

I express my sincere thanks and gratitude to our Director, BITS Pilani, Dubai Campus, **Prof. Dr. M.P. Srinivasan**,andProject Ins/ i/c **Dr. B. Vijayakumar** for their motivation, encouragement, and support to pursue my Project.

I would also like to thank my family for their constant encouragement and positivity that helped me in the successful completion of the project.

Name: **VANSH JAIN**

ID No.: **2020A7PS0079U**

**CERTIFICATE**

This is to certify that the Mid Semester Project Report entitled, **Prediction of Employee Success using Data Science** and submitted by **Vansh Jain, ID No. 2020A7PS0079U** in fulfilment of the requirement of CS F376, DESIGN PROJECT embodies the work done by her under my supervision.

Date: 22-05-2023 Signature of the Supervisor

Name: Dr. Siddhaling Urolagin

Designation: Lecturer (CS)

**BITS Pilani, Dubai Campus**

Second Semester 2022-2023

**Course Number:** CS F376

**Course Name:** DESIGN PROJECT

**Duration:** 06/02/2023 - 22/05/2023

**Date of Start:** 06/02/2023  **Date of Submission:** 22/05/2023

**Title of the Report: Prediction of Employee Success using Data Science**

**ID No. / Name of the student**: **2020A7PS0079U / VANSH JAIN**

**Discipline of Student: B.E (Hons.) Computer Science**

**Name of the Project Supervisor: Dr. Siddhaling Urolagin**

**Key Words: Employee Success Prediction, Heat maps, Classification, Employee Promotion, Attrition, correlation, Classifier,**

**Project Area: Data Science, Machine Learning**

**Abstract:**

Forecasting an employee's achievement in relation to promotions entails comprehending the elements and traits that add to a person's progress within the company's structure. This study aims to develop predictive models that can accurately forecast an employee's potential for success and identify candidates for promotion. Advanced machine learning techniques will be employed to analyse and uncover hidden patterns within the data, enabling organizations to make informed decisions about talent management and career progression. This study offers valuable insights and practical solutions for optimizing employee development and promoting a more efficient and fair promotion process within organizations. Data science and machine learning can be used to collect and analyse data on employee performance, skills, and experience. This data can be used to develop models that predict which employees are most likely to be successful in the future. These models can then be used to make more objective and accurate decisions about employee promotion. The use of data science and machine learning for employee success prediction and employee promotion has several benefits. First, it can help to reduce bias in the promotion process. Second, it can help to identify employees who are most likely to be successful in the future. Third, it can help to make the promotion process more transparent and objective.

The use of data science and machine learning for employee success prediction and employee promotion is still in its early stages. However, it has the potential to revolutionize the way that these tasks are performed. By using data science and machine learning, companies can make more informed decisions about employee promotion, which can lead to a more successful workforce.

**Signature of Student Signature of Supervisor**

**Date: 22/05/2023 Date: 22/05/2023**

**TABLE OF CONTENTS**

***ACKNOWLEDGEMENTS****2*

***CERTIFICATE FROM THE SUPERVISOR****3*

***ABSTRACT****4*

***TABLE OF CONTENTS****6*

***LIST OF FIGURES****8*

***CHAPTERS****9*

1. **Introduction**9
   1. Objective9
   2. Scope9
   3. Summary10
2. **Literature Survey**11
   1. Literature Review11
   2. Conclusion of Review16
3. **Exploratory Data Analysis**16
   1. HR Analytics: Employee Promotion Dataset.16
   2. Exploratory Data Analysis 18
      1. Unique Values19
      2. Missing/Null Values20
      3. Duplicate Values21
      4. Outliers22
      5. Correlation Matrix22
      6. Distribution Plots23
   3. Data Vizualization25
      1. is\_promoted vs gender vs age25
      2. is\_promoted vs gender vs length\_of\_service25
      3. is\_promoted vs age vs length\_of\_service26
4. **Data Cleaning And Pre-Processing**27
   1. Klib Library27
   2. SMOTE27
   3. Label Encoding28
5. Model Selection and Training29
   1. Train Test Split29
   2. Models Used29
      1. Random Forest Classifier29
      2. GridSearchCV30
      3. Decision Tree Classifier 31
      4. Logistic regression 31
      5. GaussianNB32
      6. MLP Classifier33
      7. Support Vector Classification34
   3. Model Comparision34
6. **Conclusion**36
7. **Future Works**36

*REFERENCES37*

**list of figures**

Figure 1- Before Data Cleaning (train and test merged)

Figure 2- Column Description

Figure 3- Categorical plot

Figure 4- Number of Unique Values (log values)

Figure 5- Number of null values

Figure 6- "previous\_year\_rating" test

Figure 7- "is\_promoted" class plot.

Figure 8- Outliers

Figure 9- Correlation matrix

Figure 10- Distribution plots

Figure 11- is\_promoted vs gender vs age

Figure 12- is\_promoted vs gender vs length\_of\_service.

Figure 13- is\_promoted vs age vs length\_of\_service.

Figure 14- After Data Cleaning

Figure 15- Before (Purple) and After (Orange) SMOTE

Figure 16- After label encoding

Figure 17- train test split

Figure 18- Random Forest Classifier

Figure 19- Random Forest Classifier with grid search

Figure 20- Decision Tree Classifier with grid search

Figure 21- Logistic regression

Figure 22- GaussianNB

Figure 23- MLP Classifier

Figure 24- Support Vector Classification

Figure 25 – Model Comparison

1. **Introduction**

Being able to foresee the potential achievement of an employee is a valuable resource for every organization. Organizations can enhance their hiring and training processes and foster a more nurturing work culture by identifying employees that demonstrate a high probability of success. One approach to anticipate the achievement of an employee is by examining their history of promotions. Frequently, promotion is viewed as a symbol of achievement, reflecting an employee's valuable contributions and ability to take on greater obligations.

Forecasting an employee's achievement in relation to promotions entails comprehending the elements and traits that add to a person's progress within the company's structure. Organizations can make informed decisions about talent management, development programs, and succession planning by effectively pinpointing employees who have high potential for promotion.

* 1. **Objective**
* This project aims to pinpoint the crucial elements that fuel employee performance and explore their connections to various employee outcomes.
* The purpose of this report is to present a broad understanding of the elements that lead to effective employee performance and to suggest ways to forecast such performance accurately.
* The creation of a predictive algorithm that can precisely estimate the chance of employee performance based on pertinent variables is our first goal.
* The predictive model's performance will next be assessed to establish its precision and potency as a success predictor.
* Lastly, we will assess the predictive model's results in the last step to pinpoint the most important variables that affect employee success.
  1. **Scope**

This report will analyse the employee success prediction dataset, which comprises various categories such as "employee\_id, education, department, gender, region, recruitment\_channel, no\_of\_trainings, age, previous\_year\_rating, length\_of\_service, awards\_won? avg\_training\_score, and is\_promoted". By studying how these variables are related to the outcomes of promotion, we will pinpoint the crucial elements that influence employee success and create a model that can precisely forecast success based on these elements.

* 1. **Summary**

This report delves into the utilization of data science methodologies to forecast the advancement of employees with a goal of pinpointing the variables that impact the promotion of staff members. Announcement of the ultimate promotions occurs only after evaluation, which results in a postponement of the transition to different positions. Therefore, the company requires assistance in recognizing qualified applicants at a specific stage to accelerate the entire promotion process.

This report uses data science to apply predictive modelling and predict employee success. The underlying idea is that success is indicated by promotion. Through advanced data science techniques such as statistical analysis, machine learning algorithms, and predictive modelling, the study aims to develop precise prediction models capable of forecasting an employee's chances of receiving a promotion. Exploratory data analysis (EDA) is used to find patterns and correlations between the various variables using pre-processed data of a given dataset. It easier to comprehend which elements are most important for forecasting employee performance. These patterns are found by depicting the distribution of data and find correlations between variables by using a variety of visualization techniques, including scatter plots, correlation matrices, and heatmaps.

This report aims to study the dataset thoroughly, detect patterns and correlations, and leverage data science methodologies for forecasting employee achievement. The conclusions derived from this analysis can aid organizations in developing strategies for talent management, recruitment, development, and succession planning that are based on data-driven decisions.

1. **Literature Survey**
   1. **Literature Review**
2. K. Sahinbas et al. (2022) Employee promotion is a critical decision that significantly impacts an organization. A well-designed promotion process can attract and retain top talent, enhance employee morale, and boost productivity. However, identifying the right employees for promotion can be challenging due to various factors like performance, experience, skills, and potential.

Machine learning algorithms offer a promising solution to predict employee promotions with high accuracy. These algorithms learn from historical data, enabling them to identify the most predictive factors for promotion. By leveraging this information, organizations can pinpoint employees who are most likely to succeed in new roles.

Several research studies have explored the use of machine learning algorithms for predicting employee promotions. Şahinbaş (2022) conducted a study in a large Turkish company, utilizing various machine learning algorithms. The results revealed that the Random Forest algorithm achieved the highest accuracy rate of 98%.

Şahinbaş's (2022) study involved a dataset of 1,000 employees from a Turkish company. The dataset encompassed employee information such as ID, department, region, education, gender, recruitment channel, number of trainings, age, previous year rating, length of service, awards won, average training score, and promotion status.

The data was divided into training and test sets, with the former used to train the machine learning algorithms and the latter used to evaluate their accuracy. The study incorporated the following machine learning algorithms for predicting employee promotions: logistic regression, support vector machine, random forest, decision tree, and naive Bayes.

The study found that the Random Forest algorithm achieved the highest accuracy rate of 98%, while the support vector machine algorithm ranked second with an accuracy rate of 95%. The other algorithms displayed lower accuracy rates.

These findings suggest that machine learning algorithms can predict employee promotions with a high degree of accuracy. Although the Random Forest algorithm emerged as the most accurate in this study, alternative algorithms like the support vector machine can also yield accurate predictions.

The use of machine learning algorithms for predicting employee promotions can assist organizations in making more informed promotion decisions. This, in turn, leads to a more effective promotion process that benefits both the organization and its employees.

In summary, machine learning algorithms provide a powerful tool for accurately predicting employee promotions. These algorithms can help identify employees who are likely to excel in new roles. Leveraging machine learning in promotion decisions enables organizations to enhance the effectiveness of their promotion processes and make well-informed choices.

1. F A Alqahtani and A Almaleh et al (2022) uses classification algorithms to develop predictive models for predicting whether an employee is qualified for promotion or not and identifying the most important attributes that affect employee prediction.

The research paper by Alqahtani and Almaleh uses a supervised machine learning approach to predict employee promotion. The authors used a random forest algorithm to train a model on a dataset of historical employee data. The dataset included information on employee performance, education, experience, and other factors. The model was then used to predict which employees were most likely to be promoted in the future.

The authors found that the random forest algorithm was able to predict promotion with an accuracy of 92%. This suggests that machine learning can be a valuable tool for organizations that want to improve their promotion decision-making process.

The research paper by Alqahtani and Almaleh provides evidence that machine learning can be used to predict employee promotion. The authors found that the random forest algorithm was able to predict promotion with an accuracy of 92%. This suggests that machine learning can be a valuable tool for organizations that want to improve their promotion decision-making process.

The authors of the research paper suggest that future research could focus on improving the accuracy of machine learning models for predicting employee promotion. They also suggest that future research could explore the use of machine learning to predict other employee-related outcomes, such as employee turnover and employee satisfaction.

1. A study by Jiamin Liu conducted a data-driven analysis to explore the impact of organizational status on employee promotion decisions. This literature review provides an overview of the research methodology and key findings. Examining methodologies and results provides insight into the factors that influence employee promotions and the importance of organizational status in shaping these decisions.

Liu's research used large datasets containing employee data from multiple organizations across different industries and disciplines. Statistical techniques such as regression analysis and machine learning algorithms were used to analyze the data. In this study, variables such as performance appraisal, seniority, and educational background were adjusted to explore the relationship between organizational status and employee development outcomes.

Liu's analysis shows that position within an organization plays an important role in employee promotion decisions. Employees in senior positions in a company are significantly more likely to be promoted than those in lower positions. The study also found that performance ratings interacted with position within an organization, with employees with higher ratings being more likely to be promoted regardless of their position.

Furthermore, it was found that years of service and educational background influence the relationship between position in the organization and promotion probability. Longer tenure and higher education have led to increased opportunities for promotion, especially for lower-level employees. In addition, the study highlighted industry-specific differences in the impact of organizational status on employee advancement, highlighting the impact of industry-specific factors.

Jiamin Liu's data-driven analysis provides valuable insight into the role of organizational status in talent development decisions. The study showed a positive correlation between high position within the company and potential for promotion, while also highlighting the mitigating effect of performance reviews. Furthermore, the results emphasize the intermediary role of tenure and educational attainment. These findings enhance our understanding of the factors that influence employee promotions and inform organizational promotion practices and policies.

1. The authors use two different classification techniques ANN and ANFIS for predicting employee turnover. While the application of ANN is limited, there is an advantage of using ANFIS as fuzzy logic allows for explicit knowledge through using the if-then rules. By using ANN and ANFIS, the fuzzy system can learn and understand the patterns from previously recorded datasets.

Author’s objective was to predict employee turnover in an organization using ANN and ANFIS, to see which does better from the dataset which they chose. The dataset was picked up from Kaggle and the size was around 5,000, which was then normalized to avoid any chance of overfitting. The author took 10 points for prediction.

The ANFIS method involved computing input data into clusters. Then, setting variables values of minimum, maximum, accepting ratio, rejecting ratio, and quash factor. Setting normal data value based on the minimum and maximum value. Authors then set the potential of each data value and the maximum potential of the data. The cluster centre was then identified and updates on the potential value corresponding to another data was made. Furthermore, they employed the real data within the model, set the cluster sigma, and the membership values were returned.

ANN method included assigning random weights to linkages. With the help of inputs and linkages, the NN output is determined. For the evaluation, the error is calculated, and weights are optimized using various algorithms, and this process is repeated until minimum error is recorded.

The evaluation metrics, used by the authors, were the MSE, MAE, and RMSE. For the ANN model the RMSE, MSE for training data was 0.10 and 0.01, respectively; whereas, for the ANFIS model the RMSE, MSE for training data was 0.4 and 0.23, respectively. Author’s conclusion was that the ANN model worked much better than the ANFIS model as it fit the output better for unseen data.

1. Authors proposed a ML framework to predict the retention rate of employees for an organization using a dataset. The authors used the DT, ensemble with boosted tree, KNN, and SVM. The human brain cannot handle large data and complexity of the data since the computations are very tough, thus analysis of ML algorithms must be implemented.

Making a HR based on ML model and predicting employee retention are the objectives of this paper. HR is like a store house of the data, but the managers don’t have the capabilities to do an in-depth analysis of that data. This paper implements the state-of-the-art classifiers to design best models for this prediction based on several parameters and the dataset.

Firstly, a dataset is collected, and the acquired dataset must contain past observations of different departments within a company. Next, pre-processing data is done by the authors because cleaning the data to reduce the dimensionality and the training time is vital. Analysis of the data is conducted to detect the most important features and trends within the dataset. Various ML models such as DT, SVM, KNN, etc are applied to the data for training and testing scenarios. Then, the best model is chosen in terms of accuracy to develop a predictable solution.

Employee performance measures are as important as they contribute to the commitments of employees within a company. Artificial Intelligence has proven to be helpful as it ensures transparency during the performance evaluation process.

Authors first checked the correlation between independent variables to make sure there is no multicollinearity. They found out that there is a correlation between age and length of service which was approximately 66%. Then, authors pre-processed the data to ensure that only the relevant features were fed into the model. Authors performed cleaning of the data using SMOTE, as it solved the problem of data imbalances. The irrelevant variables, employee ID and region were removed because it didn’t have any role in analysing and predicting the target variables that the authors were trying to predict.

The authors used a train-test split of 80-20. There are several different models that they tested to check if it was the best model, they were: KNN, LR, DT, RF, SVM, AB, and GB. Authors, then portrayed some charts and concluded that gender and awards won are the most important features that help in predicting employee promotion.

The evaluation metrics used by the authors are the ROC, AC, and F1S. The AC of KNN, RF, LR, SVM, DT, AB, and GB were 0.89, 0.88, 0.69, 0.89, 0.89, 0.91, and 0.93, respectively. The ROC curve was 0.62, 0.54, 0.55, 0.66, 0.65, 0.64, and 0.67, respectively. The F1S was 0.89, 0.88, 0.76, 0.89, 0.89, 0.90, 0.92, respectively.

The authors could complete their objective of developing a supervised ML model to determine promotions. HR data from MNCs were taken to predict the promotion of an employee within an organization. The results from their study indicate that there is no bias and the features recruitment did not play any role during promotion, only previous year’s ratings was the most important factor. Certain improvements that they listed were to use a larger dataset, as their dataset was very limited. If there was more data, it would lead to more optimized solutions.

1. Authors use various ML techniques to predict the employee attrition and find out the most important features that lead employees to leave companies. Prediction analysis techniques are used on HR data, then data is cleaned and pre-processed to reduce dimensionality and computation time.

The dataset was used through Kaggle which contained more than 14,00 records and all 10 features were related to the employee attrition problem. The authors used many different ML techniques such as SVM, DT, RF to predict employee attrition.

Modelling involved ML techniques that were used in experimentation and was based on ANN, DT, NB, LR, and SVM. The authors tested the different models for all the departments within a company such as the sales, technical, support, and the IT sector.

For the sales department, DT algorithm was best suited with 95% PR, 98% RC, and 97% F1S. RF model had the best evaluation metrics for the technical department with 97% PR, 97% RC, and 97% F1S. Similarly, the RF algorithm had the best performance in the support department with 98% PR, 97% RC, and 98% F1S. Moreover, the IT sector, too, had RF as its highest performing model with 99% PR, 98% RC, and 99% F1S.

All in all, the prediction for employee attrition allows organizations to identify whether the employee is going to exit the organization or not. Through the results, the authors concluded that the best model to predict employee attrition is the RF algorithm.

* 1. **Conclusion of Literature Review**

In conclusion, these research papers use machine learning methods to predict employee turnover, employee retention and employee development in organizations Authors Artificial Neural Network (ANN), Adaptive Neuro-Fuzzy Inference System (ANFIS).), Decision Trees (DT) are used to develop predictive models), support vector machines (SVM), and apply various algorithms such as ensemble methods etc.

The data are pre-processed to ensure quality and reduce dimensionality so, are mean squared error (MSE), root mean squared error (RMSE), accuracy (AC)., curves that evaluate the performance of models using metrics such as receiver operating characteristic (ROC), and F1 score (F1S). The findings show that ANN and DT models perform well in predicting employee turnover and retention, while gender and wage income are important factors in forcing employee promotion of the prophecy. This study provides valuable insights and tools for HR managers and employers to make data-driven decisions in optimizing their employees.

1. **Exploratory Data Analysis**
   1. **HR Analytics Employee Promotion Dataset**

This is a fictional data set having to dataset files train.csv and test.csv. For this project I am combining the original training and testing sets into a single data frame. This is commonly done during pre-processing and allows for consistent application of techniques such as imputation of missing values or feature engineering on the entire dataset. However, it is important to note that the united data frame should be split back into separate training and testing sets before training the ML model and making predictions.

The test and train data share identical columns with the exception of "is\_promoted", which is absent in the test data. The combination of train and test dataset gives us a new dataset of 13 columns and 78298 rows with numerical, float and object data types. The dataset was obtained through [Kaggle](https://www.kaggle.com/datasets/arashnic/hr-ana).

A screenshot of a computer

Description automatically generated

Figure 1- Before Data Cleaning (train and test merged)

**A picture containing text, screenshot, font

Description automatically generated**

Figure 2- Column Description

* 1. **Exploratory Data Analysis**

EDA involves exploring and understanding the dataset to gain insights into its characteristics. It includes tasks such as data visualization, statistical analysis, and summarization of the dataset's main features. The process of EDA involves the detection of anomalies, identifying patterns and trends, as well as spotting any missing values and outliers.

We will now try to identify and select the required features through EDA to compare against the rest of the columns. Also trying to analyse the visual patterns between different columns/features.

A picture containing text, screenshot, diagram, parallel

Description automatically generated

Figure 3- Categorical plot

* + 1. **Unique Values**

Distinct or non-repeating values that exist in a data set or a particular column/variable are known as unique values. Unique values mean the various categories or levels that a variable can have. The variable "employee\_id" has the same number of unique values as the number of rows as it is a unique id for each employee. We need to take out such columns from the dataset as they don’t affect the model and prediction in any way.

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 4- Number of Unique Values (log values)

* + 1. **Missing/Null Values**

Three columns contain missing values or null values. For the "previous\_year\_rating" and "education" columns we can first check if we can discover a pattern in the missing value. If the situation warrants, we use common sense to handle the absent information, otherwise we substitute the average value (mean for numerical fields and mode for categorical fields).

A screenshot of a computer program

Description automatically generated with medium confidence

Figure 5- Number of null values

The "education" column has missing values at random meaning that the there is no correlation/ pattern found for the null values. So, they will be filled with the mode of that column as it is a categorical column.

If someone's "length\_of\_service" is less than or equal to one, it means that they have would not have any "previous\_year\_rating" in that company. To test this, I wrote query prove it. The query below proves my hypothesis. Since they don’t have rating, the value '0' will filled instead of null.

A screen shot of a computer

Description automatically generated with low confidence

Figure 6- "previous\_year\_rating" test

"is\_promoted" is our target variable in this project. Due to combining the train and test datasets a large number of null values are present in it, since test was missing the "is\_promoted" column. For now, it will be with its mode value. It is a highly imbalanced class. To fix this imbalance, I will be applying the SMOTE technique to counter the oversampling during pre-processing.

A screen shot of a graph

Description automatically generated with low confidence

Figure 7- "is\_promoted" class plot.

* + 1. **Duplicate Values**

A duplicate value is a value that appears more than once in a dataset. Repeated values may pose issues during the analysis of data and implementation of machine learning methods. One instance when identical values can occur is: manipulate the outcome of statistical analysis in a biased manner, make it challenging to recognize patterns in the information, reduce the precision of models created through machine learning.

All the duplicate values are removed using the klib library. More about it in the pre-processing section

* + 1. **Outliers**

Outliers denote observations that deviate substantially from the rest of the data points. A variety of reasons, such as data entry errors, merging two datasets, incomplete data or natural variation can cause outliers. Outliers can influence the results of statistical analysis making it difficult to identify trends in the data also it can decrease the accuracy of machine learning models.

The Box plot is a useful tool for examining the central 50% of a dataset also referred as the interquartile range (IQR), providing clear indications of the minimum, maximum, median, and any outliers. Any data points beyond the maximum value/upper whisker/upper fence is called Outliers.

According to the plot in figure below the only outliers that would affect the model accuracy would be the "length\_of\_service" outliers. Since it the other variable have smaller number of unique values. The upper fence for "length\_of\_service" is 13, so any data points or values will be removed. The outliers present in "age" are not significant enough to cause an impact to model accuracy.

A picture containing screenshot, diagram, line, plot

Description automatically generated

Figure 8- Outliers

* + 1. **Correlation Matrix**

A correlation matrix presents a table or matrix showing the correlation coefficients among various variables in each dataset. It offers a convenient means to depict and comprehend the connections or correlations among variables. It used to identify variables that are correlated with each other, determine the strength of the relationship between variables, and identify variables that may be causing other variables to change.

A screenshot of a computer screen

Description automatically generated with low confidenceFigure 9– Correlation matrix

* age is highly overall correlated with length\_of\_service.
* avg\_training\_score is highly overall correlated with department.
  + 1. **Distribution Plots**

By contrasting the data's actual distribution with the theoretical values anticipated from a particular distribution, distribution plots visually assess the distribution of sample data. In addition to more formal hypothesis testing, distribution charts can be used to determine whether the sample data fits a particular distribution. Additionally presented are the mean, median, standard deviation, and other statistics.

A picture containing text, line, screenshot, plot

Description automatically generated

A picture containing text, line, screenshot, plot

Description automatically generated A picture containing line, plot, text, diagram

Description automatically generated A picture containing text, line, plot, font

Description automatically generated A picture containing text, line, plot, font

Description automatically generated A picture containing text, plot, line, diagram

Description automatically generated

Figure 10- Distribution plots

* 1. **Data Visualization**
     1. **is\_promoted vs gender vs age.**

Most of the employee belong to the age group of 20 to 40. Most of the promoted employees are male. While only a small number of workers are given promotions, a higher proportion of males receive promotions compared to their female counterparts. One possible reason is the imbalanced representation of females in the dataset, indicated by their lower proportion compared to males. Multiple factors can cause an imbalance in gender, e.g., variations in the demographics of the workforce or the methods used for sampling. After the age of 50 there is barely any female employee in the promoted section.

A picture containing plot, diagram, screenshot

Description automatically generated

Figure 11- is\_promoted vs gender vs age

* + 1. **is\_promoted vs gender vs length\_of\_service.**

The majority of the employees have 0 to 10 years of experience and people belonging to that group also have the majority of the promotions to themselves. People having length of service of 0 to 7 years are more likely to be promoted.

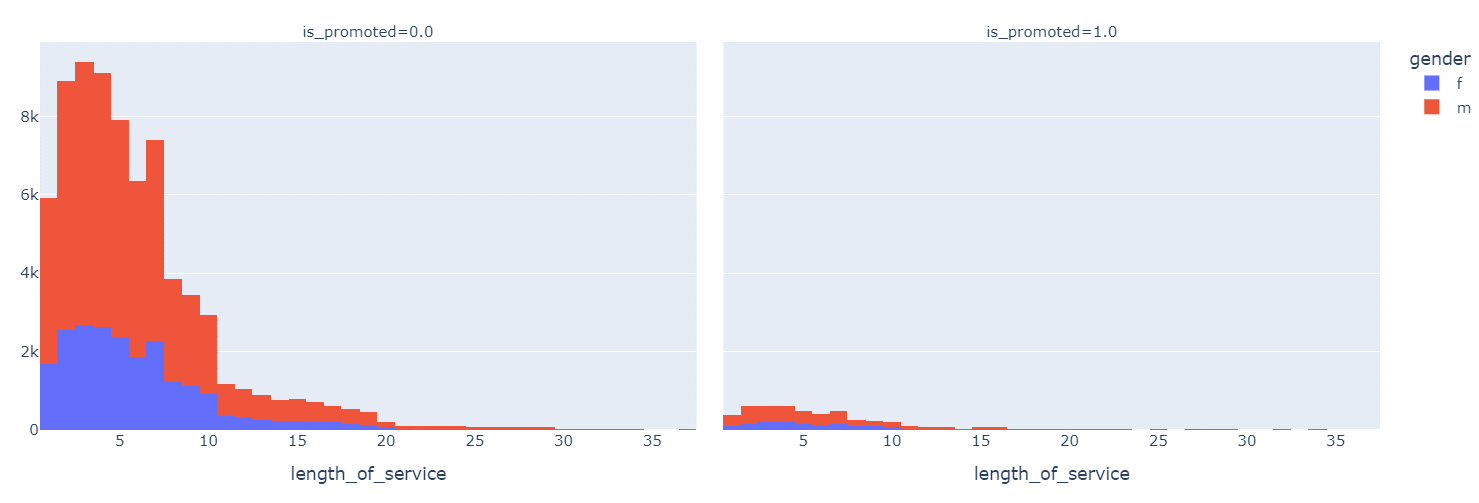


Figure 12- is\_promoted vs gender vs length\_of\_service.

* + 1. **is\_promoted vs age vs length\_of\_service.**

By examining the strip plot, it becomes apparent that there is a correlation of a favourable nature between the length of service and age. Usually, as employees stay longer in their job, the range of ages they belong to tends to expand or widen. This discovery indicates that the seniority of employees in the company is positively correlated with their age. This association corresponds with rational thinking. Less-experienced workers who are younger have had fewer opportunities to dedicate themselves to their careers and amass knowledge within the company. On the other hand, senior workers probably have a more extended history with the organization, suggesting greater length of service.

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 13- is\_promoted vs age vs length\_of\_service.

1. **Data Cleaning and Pre-processing**

The data cleaning involves checking the dataset for missing values, duplicate entries, and incorrect data types. Due to the merging of the datasets the number of missing values and duplicates had increased. To resolve these imperfections, different methods were applied.

* 1. **Klib Library**

Klib is a Python library that presents a diverse range of tools for conducting data exploration and visualization. The goal is to make the beginning stages of data analysis easier and more efficient by presenting rapid and easy-to-use methods for comprehending and presenting sets of data.

The data\_cleaning() function in Klib is used to clean data. It can be used to remove missing values, convert data types, and fix errors. The function takes a data frame as input and returns a cleaned data frame. Another is useful feature is that it removes any duplicate rows or columns present in dataset. All this cleaning greatly reduces the memory size of the data in turn reducing the time required to train a model using it.

A screenshot of a computer

Description automatically generated

Figure 14- After Data Cleaning

* 1. **SMOTE (Synthetic Minority Oversampling Technique)**

The missing values in "education" was resolved by filling them with the mode value. For "previous\_year\_rating", 0 was filled in place of null since people with less than or equal one year would not have a previous rating. "is\_promoted" column had a high class imbalance so filling it with the mode value doesn’t resolve the issue. Class imbalance occurs when a dataset contains an uneven distribution of a particular class. This means that the quantity of data points in the majority class (negative class) is considerably greater than that in the minority class (positive class). Failing to address imbalanced data may lead to decreased effectiveness of the classifier model. Many of the forecasts are expected to align with the larger group, disregarding the features of the smaller group as insignificant and disregarding them in the data. The model will have a significant bias as an outcome of this.

There are two main approaches for this issue: under sampling, which involves reducing the size of the dataset, and oversampling, which involves increasing the size of the dataset. Oversampling is more favoured since under sampling involves eliminating data instances which could potentially contain valuable information.

SMOTE is a method for increasing the representation of the underrepresented class through the creation of artificially generated samples. This algorithm offers a solution to the issue of overfitting caused by utilizing random oversampling. The technique centres on utilizing the feature space for the creation of fresh examples by means of interpolation between closely positioned positive instances.

A picture containing screenshot, text, rectangle, plot

Description automatically generated

Figure 15 – Before (Purple) and After (Orange) SMOTE

* 1. **Label Encoding**

Label Encoding is a method for converting category columns into numerical ones so that machine learning models, which can only handle numerical data, can fit them. Each unique category in a categorical variable is assigned a different integer label.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 16- After label encoding

1. **Model Selection and Training**

Predictive modelling is a process that involves utilizing data and statistical techniques to develop models that make predictions or estimations concerning forthcoming results through analysing past data trends. It holds great importance in numerous domains such as commerce, economics, medical care, and advertising, etc.

* 1. **Train Test Split**

Any data before used for modelling must be split into a train set to train and teach the model and a test set for evaluating the model's performance and accuracy. The train-test split is typically performed randomly, ensuring that the data points in the training set and the test set are representative of the overall dataset. By splitting the data the model's performance can be evaluated on unseen data. The proportion of the split can vary depending on the size and nature of the dataset.

I am taking train as 80% and test to be 20%.

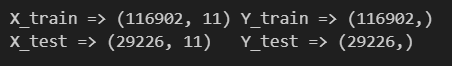


Figure 17- train test split

* 1. **Models Used**
     1. **Random Forest Classifier**

The Random Forest Classifier is a type of algorithm used for supervised learning that utilizes ensemble learning to produce a model based on a set of decision trees. This algorithm holds significant strength in its capability to perform well in both classification and regression tasks.

The fundamental concept involves generating numerous decision trees, which are trained on various random subsets of the training dataset. The final prediction is obtained by taking the average of the forecasts generated by each individual tree. This technique aids in mitigating overfitting, a common issue associated with decision trees.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 18- Random Forest Classifier

* + 1. **GridSearchCV**

Grid search involves thoroughly exploring a range of hyperparameter values in order to discover the optimal combination of parameters that yields the highest performance on a specific dataset. Whilst it may be computationally costly, using a brute-force technique is a trustworthy method for obtaining optimal hyperparameters for a specific model.

Before conducting a grid search, it is essential to establish a framework of potential hyperparameter values that will be explored. You will train and test a model for every hyperparameter combination specified in the grid using the training and test sets.

A screenshot of a computer

Description automatically generated

Figure 19- Random Forest Classifier with grid search

* + 1. **Decision Tree Classifier**

The Decision Tree Classifier is a form of supervised learning technique that can be applied in carrying out both regression and classification activities. A top-down approach is employed by a tree-based algorithm to formulate predictions. The primary concept of the classifier is to repeatedly divide the data into more compact subsets until every subset is homogenous. A pure subset refers to a subset containing data points exclusively of one class.

The splitting process utilizes a decision tree as its foundation. A decision tree is a graphical depiction of the process of data splitting. The decision tree consists of nodes that indicate the split points and leaves that signify the pure subsets.

The classifier begins making predictions by first locating the root of the decision tree, and then navigating through the various branches until a relevant leaf is found that corresponds with the particular data point under consideration. The category of the data point is subsequently designated as the same category as that of the leaf.

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 20- Decision Tree Classifier with grid search

* + 1. **Logistic Regression**

Logistic regression is a statistical tool that can be utilized to estimate the chance of a categorical event occurring based on certain independent factors. The binary dependent variable used in logistic regression has only two possible values, such as "failure" or "success," "yes" or "no," or 0 or 1. Independent variables in this model can either be continuous or categorical.

Logistic regression analysis differs from linear regression analysis as it is utilized to forecast a categorical result, rather than a continuous one. Linear regression employs a mathematical equation of a straight line to represent the connection between the autonomous elements and the reliant element. In contrast, logistic regression employs a logistic equation to depict how the independent variables are related to the dependent variable.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 21- Logistic regression

* + 1. **GaussianNB**

Gaussian Naive Bayes (GaussianNB) is a supervised learning algorithm that is used for classification tasks. It is a simple algorithm that is based on the Bayes theorem. The Bayes theorem, which is a mathematical equation, can be utilized to determine the likelihood of an occurrence of an event, taking into account the probability of other related events. The probability of a data point belonging to a specific class is determined in GaussianNB by utilizing the Bayes theorem and considering the values of its characteristics.

GaussianNB supposes that the features of every category follow a typical distribution. This indicates that the characteristics follow a normal distribution, with distinct mean and standard deviation values for every group, forming a curve that resembles a bell shape.

GaussianNB uses probability calculations for each data point to predict which class it belongs to. The class with the highest probability is selected to assign the data point.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 22- GaussianNB

* + 1. **MLP Classifier**

MLP Classifier, or Multi-layer Perceptron Classifier is a feedforward neural network that consists of multiple layers of nodes (neurons) with nonlinear activation functions. MLP Classifier is a supervised learning algorithm that learns a function f(⋅):Rm→Rn by training on a dataset, where m is the number of dimensions for input and n is the number of dimensions for output. Given a set of features X = x1,x2,...,xm and a target y, it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers.

A screenshot of a computer

Description automatically generated

Figure 23- MLP Classifier

* + 1. **Support Vector Classification**

Support Vector Classification (SVC) is a method of supervised learning employed to carry out classification tasks. This belongs to the category of SVMs. SVMs refer to a category of tools employed for accomplishing classification and regression objectives.

SVC operates by identifying a hyperplane that effectively partitions the data points of two distinct classes. The hyperplane is a geometric boundary, which can either take the form of a straight line or flattened surface, that partitions the dataset into two distinct groups in such a way that all observations within each group are members of the same class, while those outside of the hyperplane belong to the opposite class.

SVC identifies the hyperplane that achieves the maximum distance between the two classes, known as the margin. The margin refers to the measurement of the space from the hyperplane to the data points that are closest in each category.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 24- Support Vector Classification

* 1. **Model Comparison**

Comparing all the models after testing their performance, the Random Forest Classifier has the highest accuracy, followed by the Decision Tree Classifier. The Logistic Regression model has the lowest accuracy. It should be noted that the precision of a model may fluctuate based on the dataset used for its training. It is crucial to assess a model's precision across various datasets prior to settling on a model choice.

The most suitable model to utilize will be determined by the particular requirements that are needed. If your goal is to find a highly accurate model, consider opting for either the Random Forest or Decision Tree Classifier. If your intention is to find a straightforward model to comprehend, the Logistic Regression model could be a suitable option. If you're seeking a model capable of managing intricate relationships among characteristics, opting for the C-Support Vector Classification model might be a good decision.

It's important to consider other factors such as computational complexity, interpretability, and the specific requirements of the problem before finalizing the choice of the model.

**A picture containing text, screenshot, software, multimedia software

Description automatically generated**

Figure 25 – Model Comparison

1. **Conclusion**

The objective of this project is to predict the promotion of employees based on certain characteristics. Being promoted generally means that a person has worked hard and received results. In different words that person has achieved success. The analysis covers stages such as sorting out the data, data cleaning to exploration to model training and evaluation before going to the most important part necessary for answering the questions introduced at the beginning, which is the interpretability or comprehensibility of the model. The findings indicated that gender, previous\_year\_rating, avg\_training\_score,number\_of\_trainings has an impact on the effectiveness of the prediction.

1. **Future Works**

* Need more data to improve prediction accuracy. Additional data pre-processing needed to enhance the model's performance.
* Data Augmentation: If the dataset is limited in size, data augmentation techniques can be employed to generate additional training samples. This can help improve the models' generalization and overall performance.
* Advanced Neural Network Architectures: Exploring more advanced architectures, such as deep neural networks, convolutional neural networks (CNNs), or recurrent neural networks (RNNs), could potentially yield better results, especially if the data has specific structural properties or sequential dependencies.
* Enhancing the user interface: The project could be improved by developing a more user-friendly interface, making it easier for non-technical users to access.

**References**

1. K. Şahinbaş, "Employee Promotion Prediction by using Machine Learning Algorithms for Imbalanced Dataset," 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Istanbul, Turkey, 2022, pp. 1-5, doi: 10.1109/ICMI55296.2022.9873744.
2. F. A. Alqahtani and A. Almaleh, "Analysis and Prediction of Employee Promotions Using Machine Learning," 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2022, pp. 01-09, doi: 10.1109/DSIT55514.2022.9943959.
3. J. Liu, T. Wang, J. Li, J. Huang, F. Yao, and R. He, "A Data-driven Analysis of Employee Promotion: The Role of the Position of Organization," 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 4056-4062, doi: 10.1109/SMC.2019.8914449.
4. U. Soni, N. Singh, Y. Swami, and P. Deshwal, "A Comparison Study between ANN and ANFIS for the Prediction of Employee Turnover in an Organization," 2018 International Conference on Computing, Power, and Communication Technologies (GUCON), Greater Noida, India, 2018, pp. 203-206, doi: 10.1109/GUCON.2018.8674886.
5. B. Kaur and A. Dogra, "A Machine Learning Model for Predicting Employees Retention: An Initiative towards HR through Machine," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 653-657, doi: 10.1109/PDGC56933.2022.10053249.
6. Jain, Praphula Kumar, et al. “Explaining and Predicting Employees’ Attrition: A Machine Learning Approach.” *SN Applied Sciences*, vol. 2, no. 4, 2020, https://doi.org/10.1007/s42452-020-2519-4.