Employee Performance Analysis

Team Members : **Anand P & Anthony Al Alam**

Department : Artificial Intelligence Systems

University: **EPITA, l'école des ingénieurs en intelligence informatique**

Date of Submission : 30-07-2022

Project Summary

Requirement

INX Future Inc, (referred as INX), is one of the leading data analytics and automation solutions provider with over 15 years of global business presence. INX is consistently rated as top 20 best employers past 5 years. INX human resource policies are considered as employee friendly and widely perceived as best practices in the industry. Recent years, the employee performance indexes are not healthy, and this is becoming a growing concern among the top management. There have been increased escalations on service delivery and client satisfaction levels came down by 8 percentage points.

CEO, Mr. Brain, knows the issues but concerned to take any actions in penalizing non-performing employees as this would affect the employee morale of all the employees in general and may further reduce the performance. Also, the market perception best employer and thereby attracting best talents to join the company. Mr. Brain decided to initiate a data science project, which analyses the current employee data and find the core underlying causes of this performance issues

The following insights are done in this project.

1. Department wise performances

2. Top 3 Important Factors effecting employee performance

3. A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees.

4. Recommendations to improve the employee performance based on insights from analysis.

The entire project is done in **JUPITER** notebook by using Python language.

5: Model to deploy for Industrilisation are done in and python packages have been uploaded accordingly (src>app>

- inference.py
- preprocessing.py
- train_model.py

6.Integrated MFLOW and retrained the MODEL to Predict , tracking model metrics as well

Analysis

It is a classification problem. The data that is being provided consists of categorical fields and numerical fields.

The categorical fields are

Gender, EducationBackground, MaritalStatus, EmpDepartment, EmpJobRole, Business Travel Frequency, Overtime, Attrition.

These values are nominal, ordinal, ratio or interval.

The numerical fields are

Age, DistanceFromHome, EmpEducationLevel, EmpEnvironmentSatisfaction, EmpHourlyRate, EmpJobInvolvement, EmpJobLevel, EmpJobSatisfaction, NumCompaniesWorked, EmpLastSalaryHikePercent, EmpRelationshipSatisfaction, TotalWorkExperienceInYears, TrainingTimesLastYear, EmpWorkLifeBalance, ExperienceYearsAtThisCompany, ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, PerformanceRating.

These values are either discrete or continuous.

The target variable **'PerformanceRating'** is ordinal.

Step 1: **Perform Exploratory Data Analysis (EDA).**

It includes

1) Checking the datatypes of data.

2) Finding the names of the columns present in the data, shape of the data, information of the data and describing the data.

3) Checking for the null values if they are present in the data and remove them.

Step 2: **Perform Visualization (Graphical Representation) in order to carry out detailed analysis.**

1) Here, department wise performance analysis is carried out.

2) Also each and every factor related to employee performance is being analyzed.

The details of visualization are described in the Summary and Insights.

Step 3: define X and y variables.

Here X represents input variables and y represents output variable.

Step 4: Using Train-Test split, scaling, and standard scaling.

Train-Test split splits arrays or matrices into random train and test subsets.

Scale standardizes a dataset along any axis.

Standard Scaler standardizes features by removing the mean and scaling to unit variance.

Step 5: Machine Learning Algorithm to predict the employee performance.

Algorithms used are

- Random Forest Classifier

Steps for designing the machine learning algorithms

1. Import the required packages.
2. Define and train the model.
3. Predict the model.
4. Calculate the accuracy score, precision score, recall score and F1 score.
5. Display the confusion matrix and crosstab.
6. Display the classification report.

Other techniques used include:

1. **Feature Engineering** in Random Forest Classifier

Steps:

A. Import the required package.
B. Sort the values as per the correlation with respect to Performance Rating.
C. Define X(input) and y(output) variables.
D. Use train-test split to divide test and train data.
E. Define the model.
F. Predict the model.
G. Display the confusion matrix and crosstab.
H. Calculate the accuracy score, precision score, recall score and F1 score.
I. Display the classification report.

2. **Randomized Search Cross Validation** (CV) in Random Forest Classifier

Steps:

A. Import the required package.
B. Use train-test split and standard scaler.
C. Define and train the model.
D. Find best_score_ and best_params_ values.
E. Predict the model.
F. Display the confusion matrix and crosstab.
G. Calculate the accuracy score, precision score, recall score and F1 score.
H. Display the classification report.

A correlation matrix is being created to understand the relation of all the fields with respect to Performance Rating.

The factors which are **positively correlated** with Performance Rating are **Environment Satisfaction, Last Salary Hike Percent and Work Life Balance**.

The factors which are **negatively correlated** with Performance Rating are **Years Since Last Promotion, Experience Years In Current Role, Years With Current Manager and Experience Years At This Company.**

Also, a technique called Label Encoder is used to convert the categorical data into numerical data so that the predictive models can understand the data. The fields which are converted to numericals using Label Encoder are 1) **EmpNumber**, 2) **Gender**, 3) **EducationBackground**, 4) **MaritalStatus**, 5) **EmpDepartment**, 6) **EmpJobRole**, 7) **BusinessTravelFrequency**, 8) **Overtime**, 9) **Attrition**.

Summary

In this project, we try to figure out which department has performed well, factors which affect employee performance and train a model using machine learning algorithm to predict the Performance Rating.

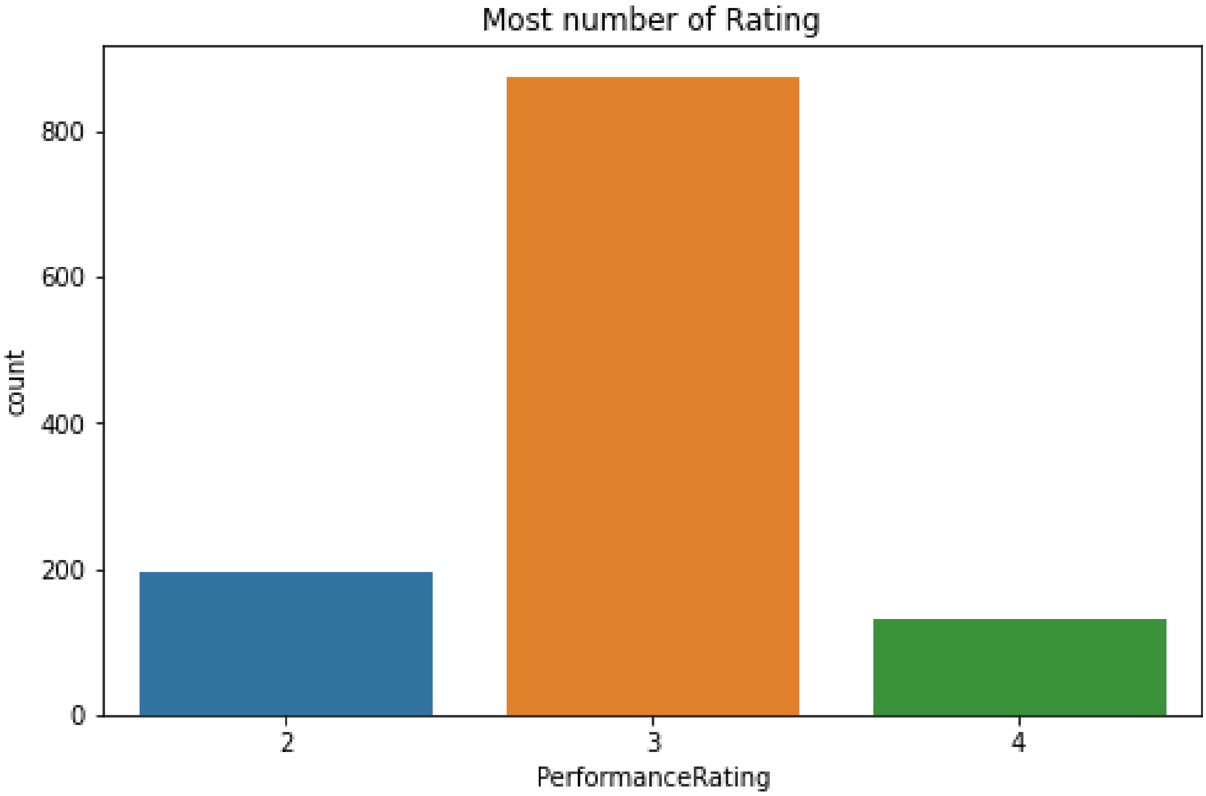
We also analyze the data and give recommendation to improve the employee’s performance.

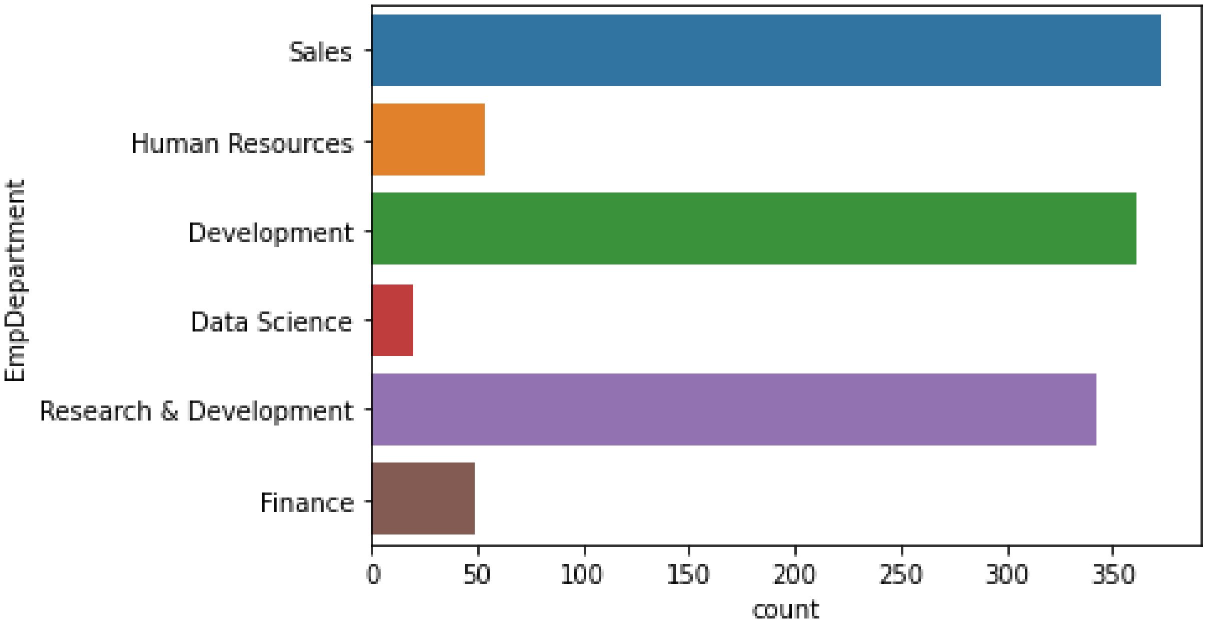
1. Department wise performances

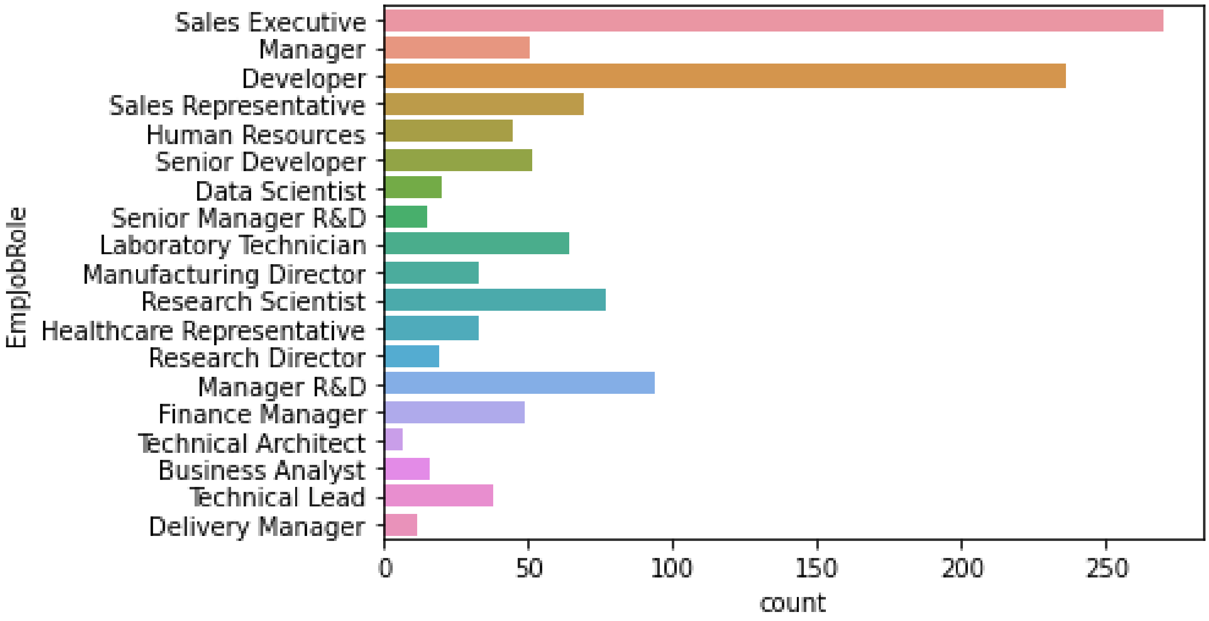
By using the field called Performance Rating and finding the mean of the values for all the departments, we can conclude that department which has the highest average performance rating is '**Development**' and the department which has the lowest performance rating is '**Finance**'.

The following are the average Performance Ratings of each Department:
Data Science --> 3.050000
Development --> 3.085873
Finance --> 2.775510
Human Resources --> 2.925926

Research & Development --> 2.921283
Sales --> 2.860590



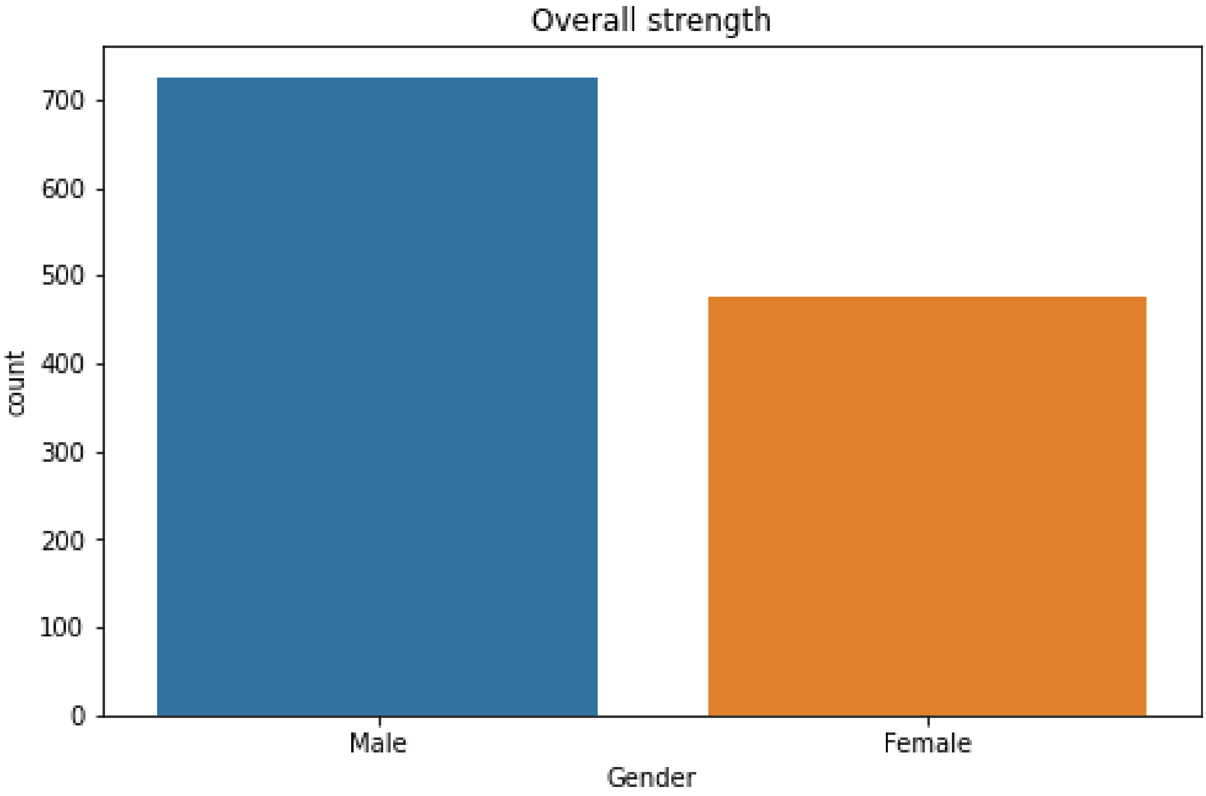


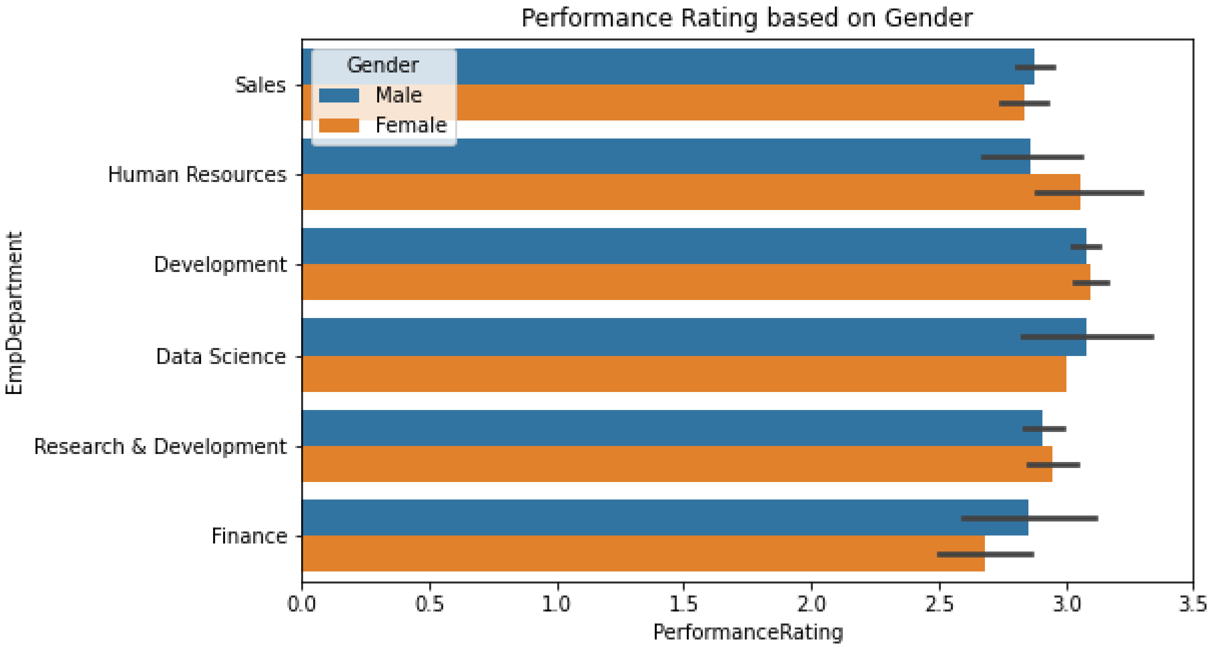


We can see majority of the employee's are from the

- Sales Executive department
- Developer

PERFOMANCE BASED ON GENDER

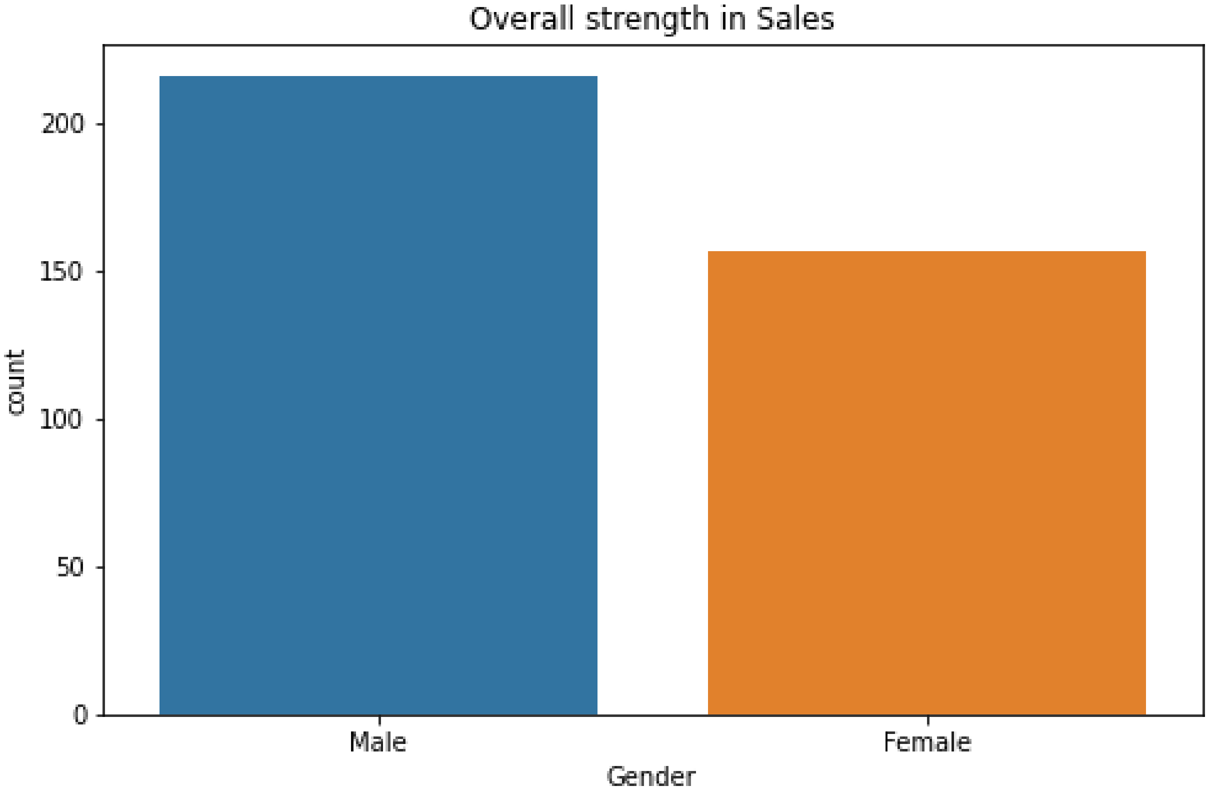


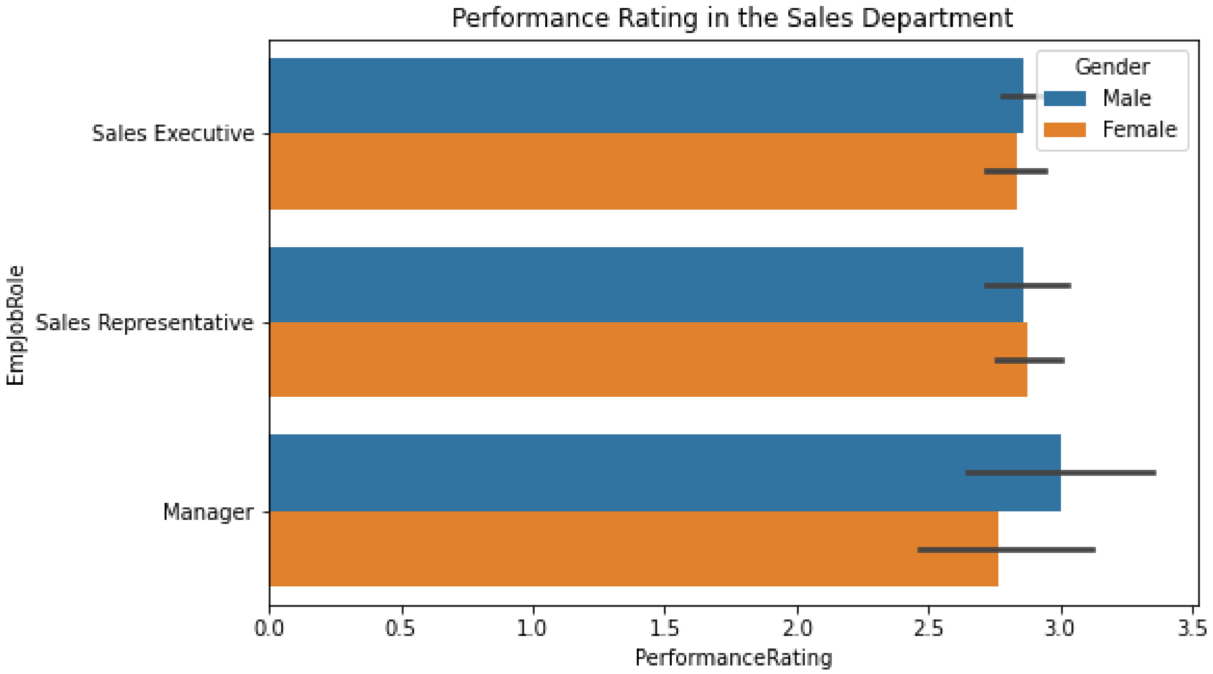


We can see that ,

- **Male have higher** performance rating in the sales department
- **Females have higher** performance rating in the HR department
- **Male & Female** have almost the same performance rating in the development team
- **Male have higher** performance rating in the data science department
- **Male & Female** have almost equal performance in the R&D team
- **Male has outperformed female** in the finance department

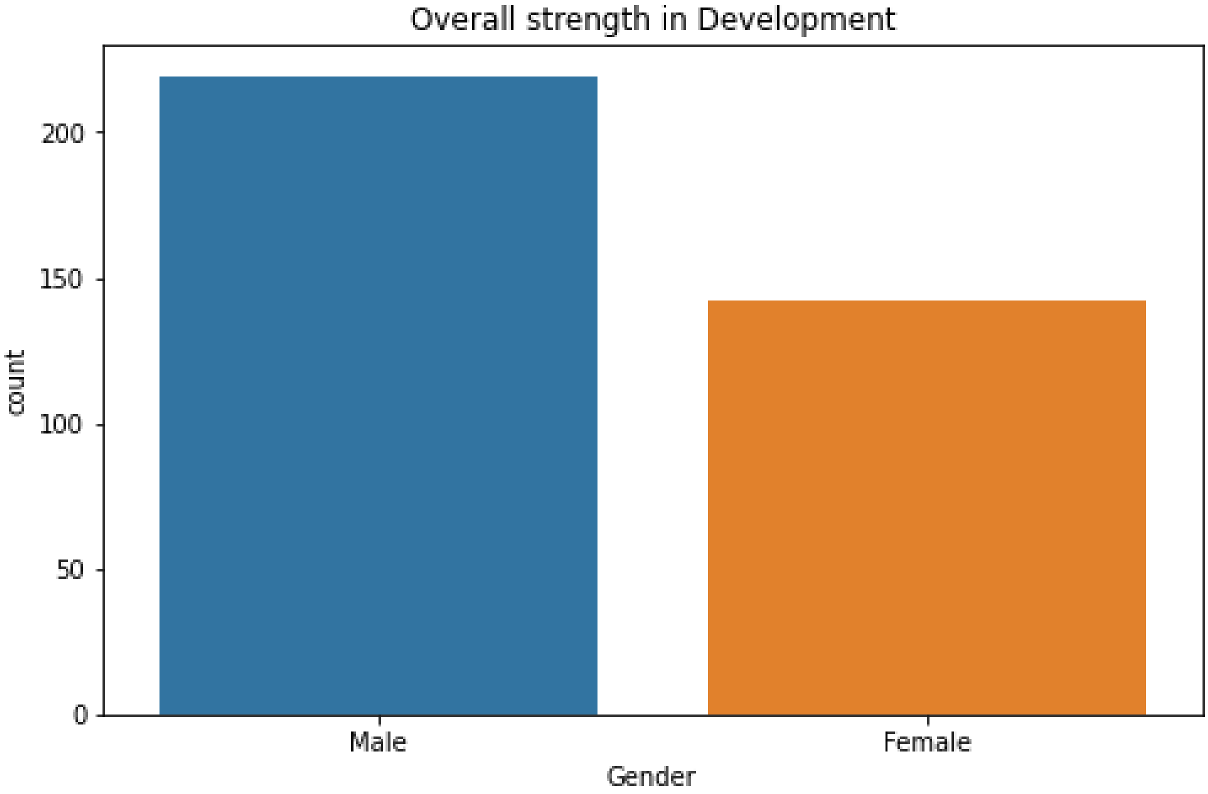
DEPARTMENT WISE SALES

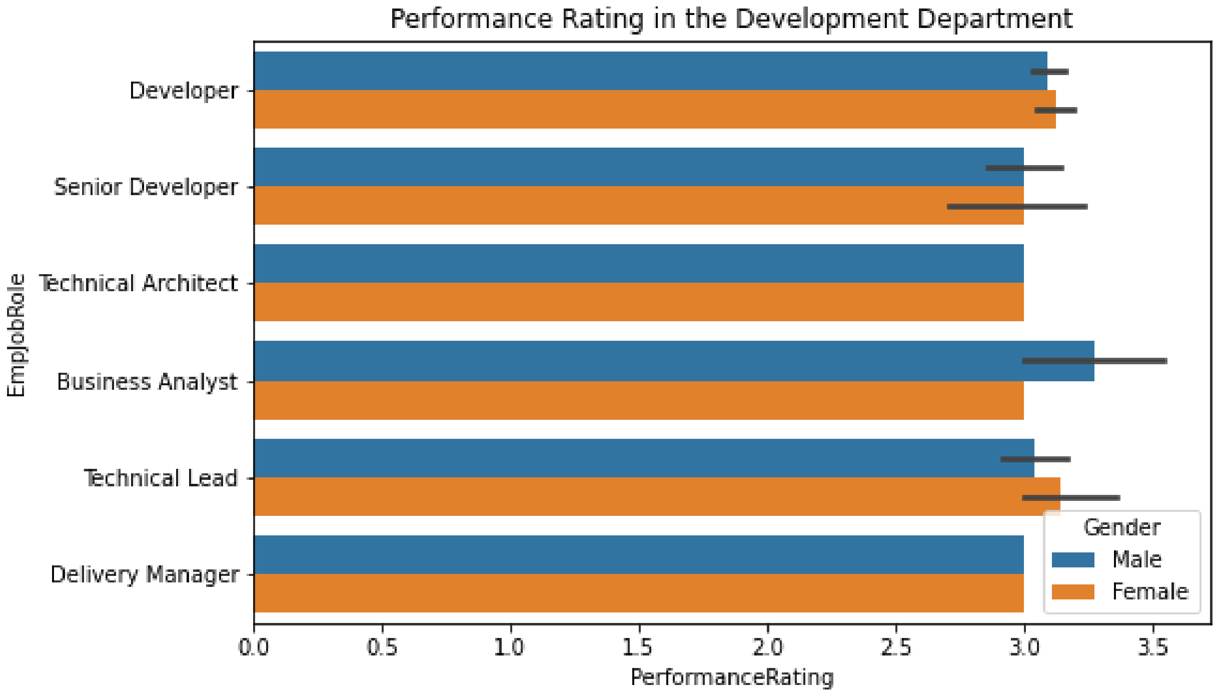




In terms of manager , male has a higher performance rating than female , other than that , the rest of the roles share the same performance rating

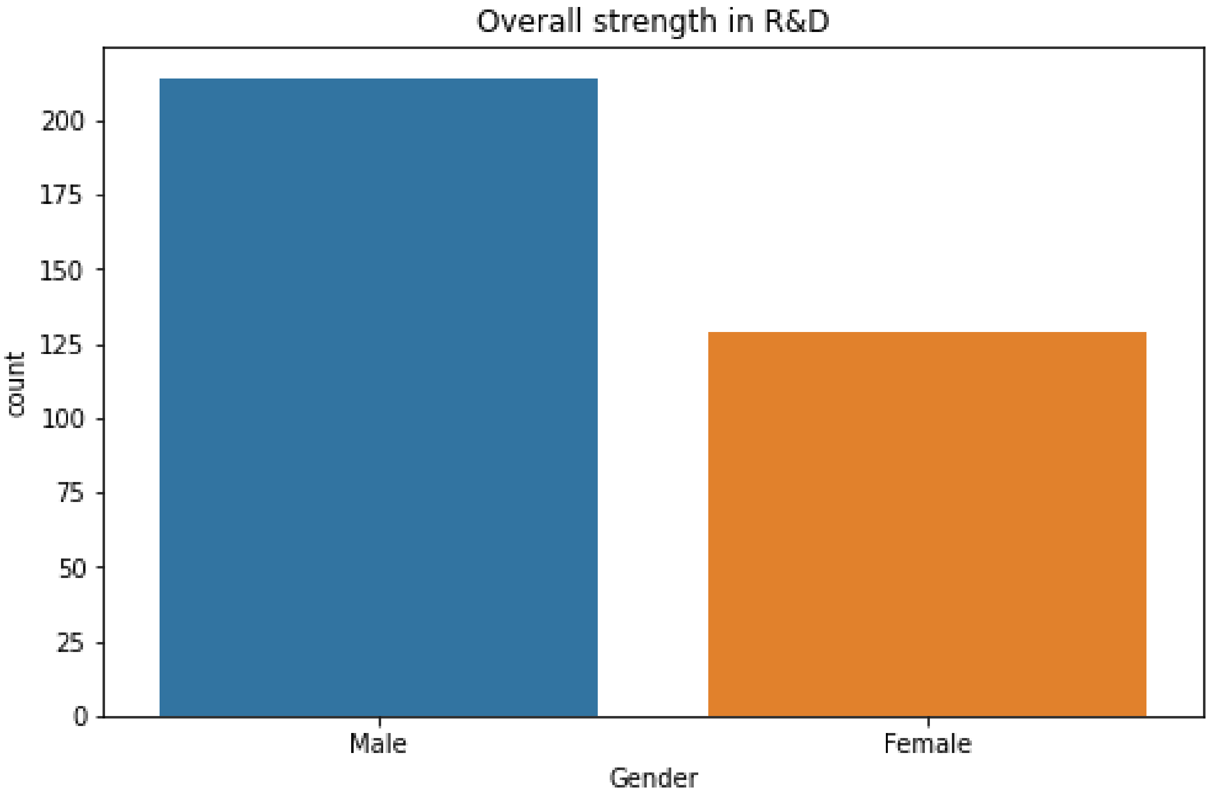
DEPARTMENT WISE DEVELOPER

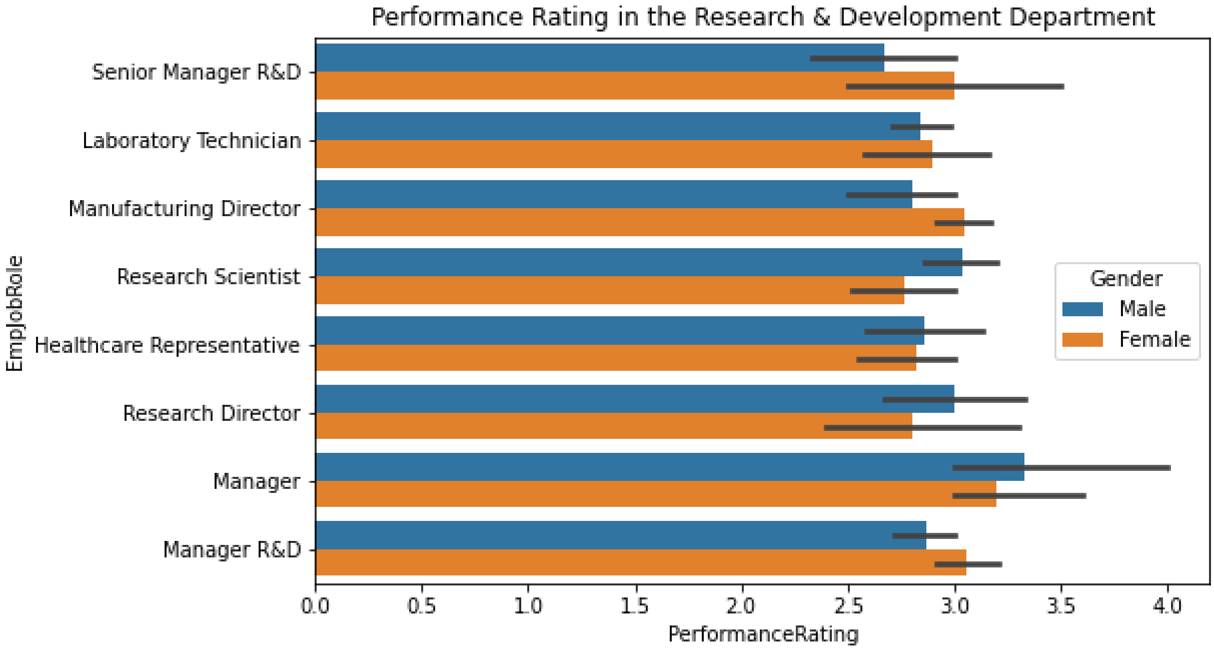




Performance rating the development team is almost the same with an average performance rating of 3 , meaning that all of them are satisfied with the job

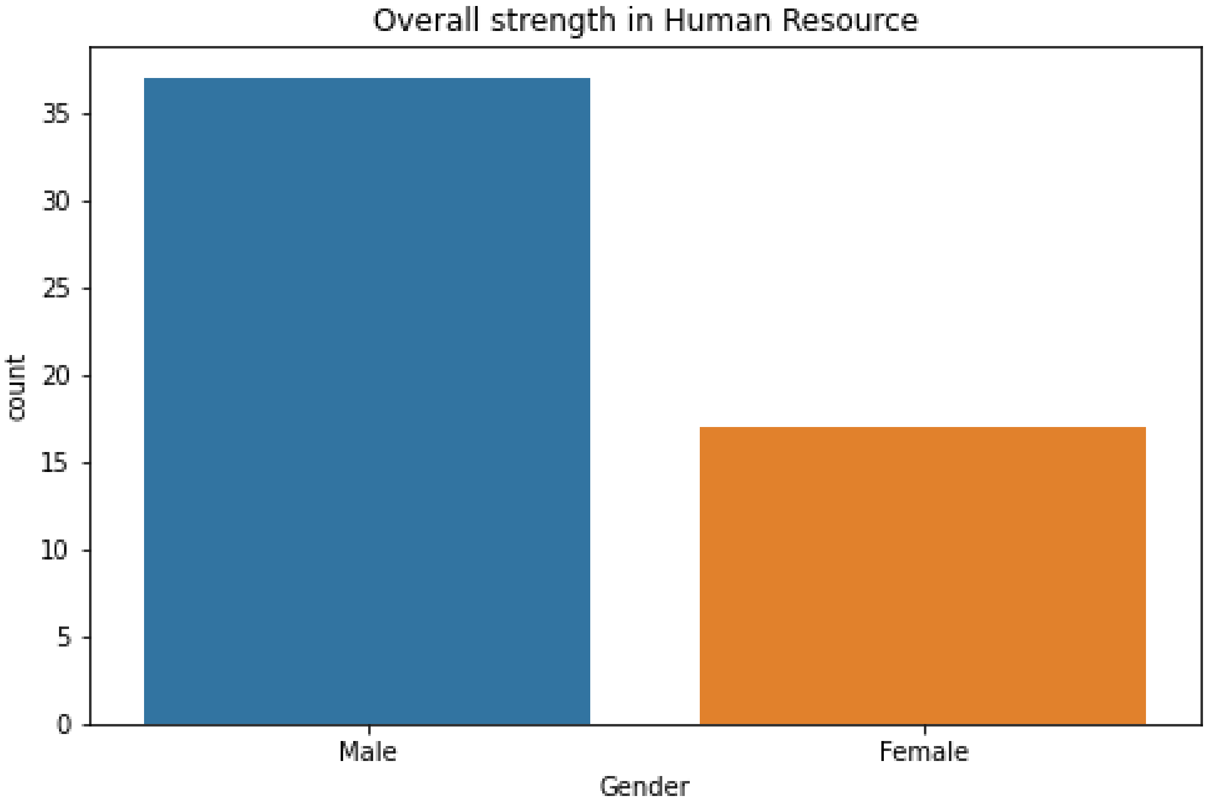
DEPARTMENT WISE R&D

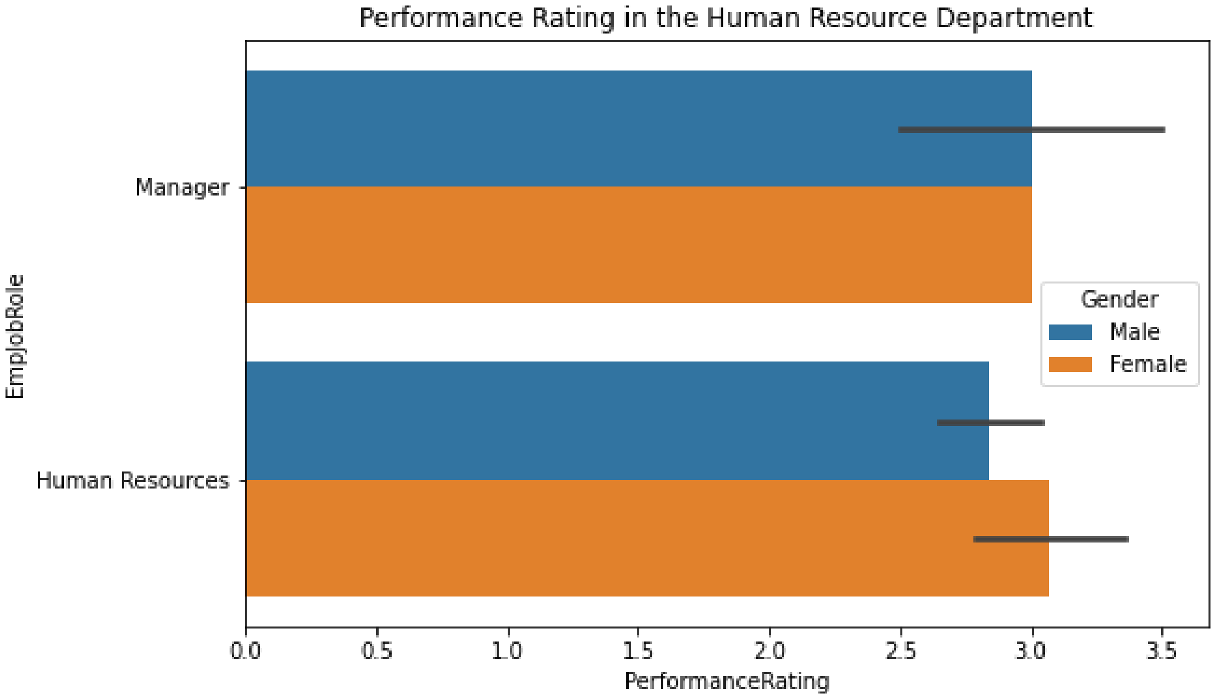




Manager has a higher performance rating , compared to the rest of the roles in the R&D team and the rest of the roles share a similar performance rating which is about 2.5 – 3

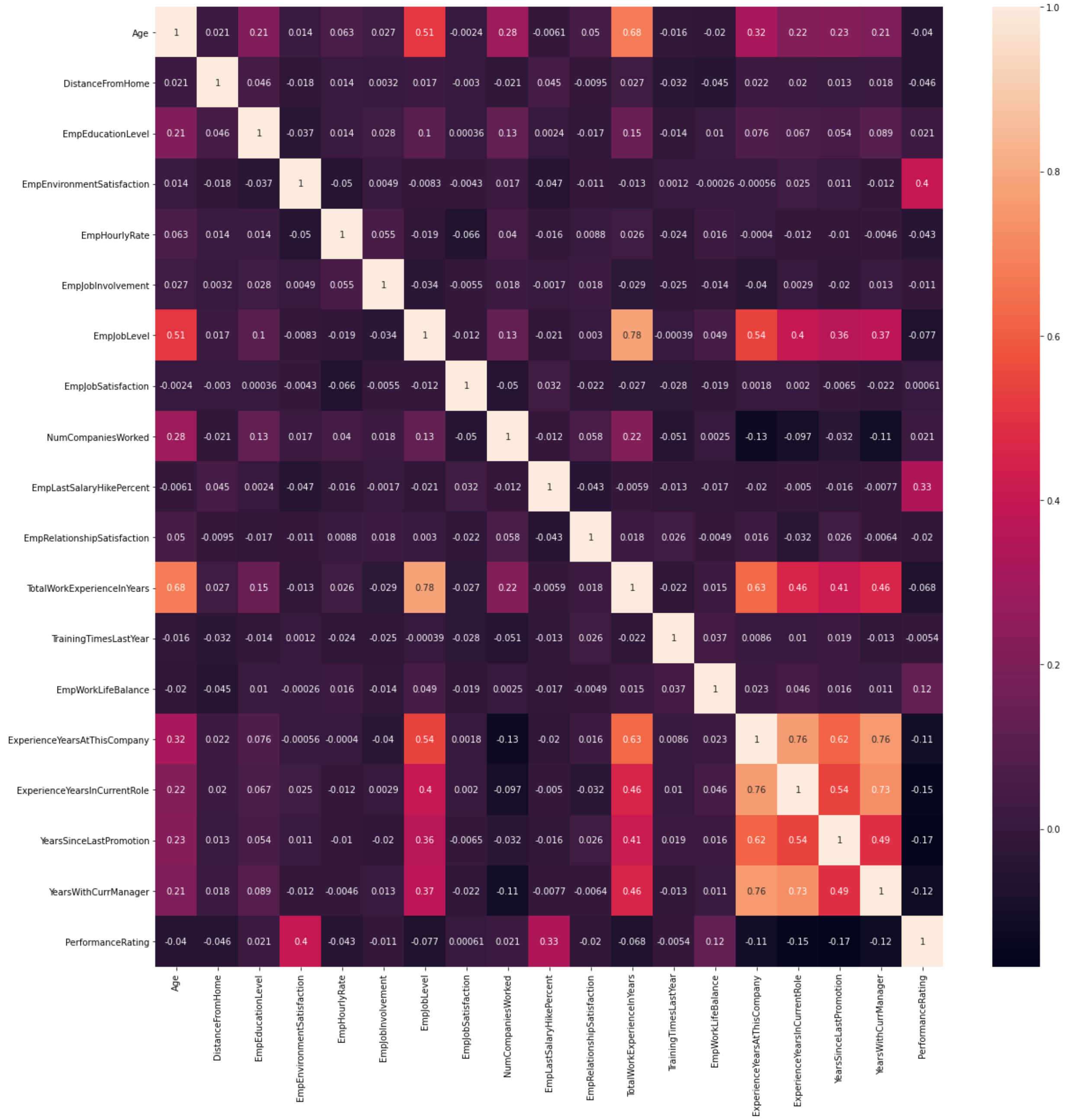
DEPARTMENT WISE HUMAN RESOURCES





Females compartively have higher performance rating that male in terms of HR as well as well as manager

FACTORS AFFECTING THE PERFORMANCE RATING



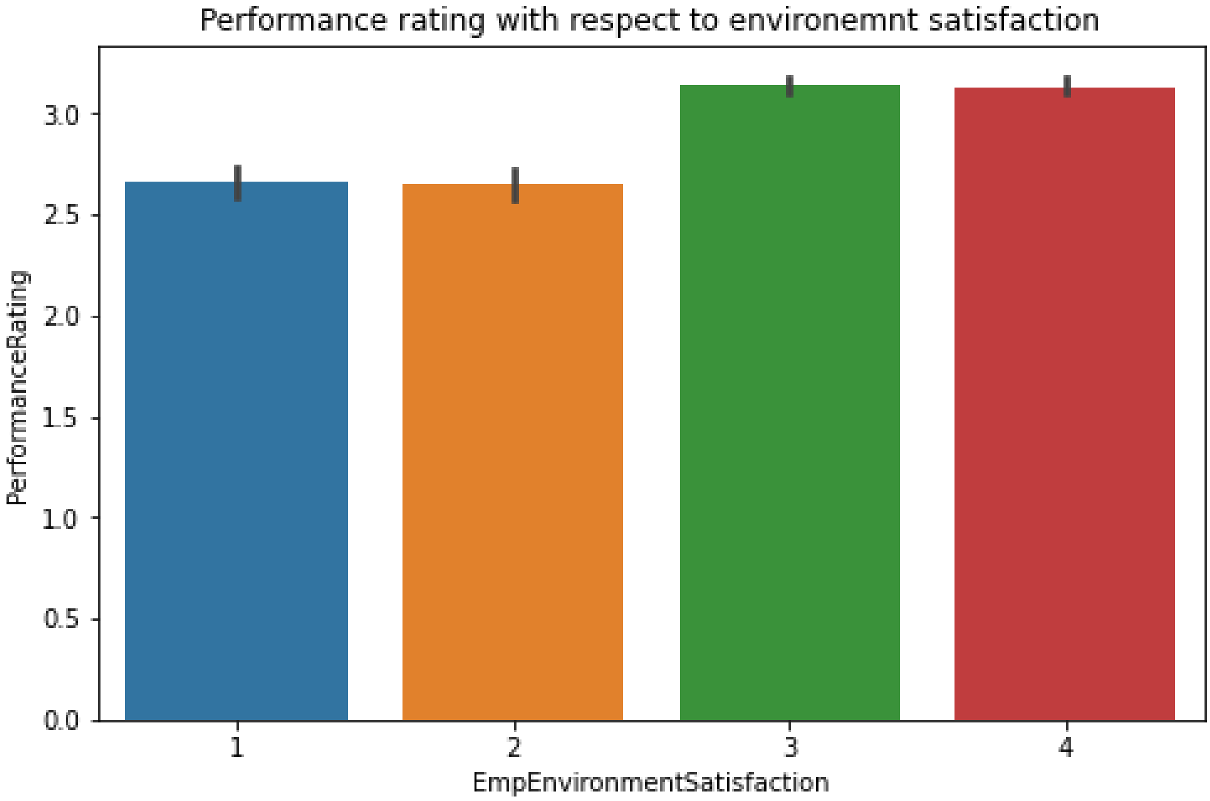
	Age	DistanceFromHome	EmpEducationLevel	EmpEnvironmentSatisfaction	EmpHourlyRate	EmpJobInvolvement	EmpJobLevel	EmpJobSatisfaction	NumCompaniesWorked	EmpLastSalaryHikePercent	EmpRelationshipSatisfaction	TotalWorkExperienceInYears	TrainingTimesLastYear	EmpWorkLifeBalance	ExperienceYearsAtThisCompany	ExperienceYearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	PerformanceRating
Age	1	0.021	0.21	0.014	0.063	0.027	0.51	-0.0024	0.28	-0.0061	0.05	0.68	-0.016	-0.02	0.32	0.22	0.23	0.21	-0.04
DistanceFromHome	0.021	1	0.046	-0.018	0.014	0.0032	0.017	-0.003	-0.021	0.045	-0.0095	0.027	-0.032	-0.045	0.022	0.02	0.013	0.018	-0.046
EmpEducationLevel	0.21	0.046	1	-0.037	0.014	0.028	0.1	0.00036	0.13	0.0024	-0.017	0.15	-0.014	0.01	0.076	0.067	0.054	0.089	0.021
EmpEnvironmentSatisfaction	0.014	-0.018	-0.037	1	-0.05	0.0049	-0.0083	-0.0043	0.017	-0.047	-0.011	-0.013	0.0012	-0.00026	-0.00056	0.025	0.011	-0.012	0.4
EmpHourlyRate	0.063	0.014	0.014	-0.05	1	0.055	-0.019	-0.066	0.04	-0.016	0.0088	0.026	-0.024	0.016	-0.0004	-0.012	-0.01	-0.0046	-0.043
EmpJobInvolvement	0.027	0.0032	0.028	0.0049	0.055	1	-0.034	-0.0055	0.018	-0.0017	0.018	-0.029	-0.025	-0.014	-0.04	0.0029	-0.02	0.013	-0.011
EmpJobLevel	0.51	0.017	0.1	-0.0083	-0.019	-0.034	1	-0.012	0.13	-0.021	0.003	0.78	-0.00039	0.049	0.54	0.4	0.36	0.37	-0.077
EmpJobSatisfaction	-0.0024	-0.003	0.00036	-0.0043	-0.066	-0.0055	-0.012	1	-0.05	0.032	-0.022	-0.027	-0.028	-0.019	0.0018	0.002	-0.0065	-0.022	0.00061
NumCompaniesWorked	0.28	-0.021	0.13	0.017	0.04	0.018	0.13	-0.05	1	-0.012	0.058	0.22	-0.051	0.0025	-0.13	-0.097	-0.032	-0.11	0.021
EmpLastSalaryHikePercent	-0.0061	0.045	0.0024	-0.047	-0.016	-0.0017	-0.021	0.032	-0.012	1	-0.043	-0.0059	-0.013	-0.017	-0.02	-0.005	-0.016	-0.0077	0.33
EmpRelationshipSatisfaction	0.05	-0.0095	-0.017	-0.011	0.0088	0.018	0.003	-0.022	0.058	-0.043	1	0.018	0.026	-0.0049	0.016	-0.032	0.026	-0.0064	-0.02
TotalWorkExperienceInYears	0.68	0.027	0.15	-0.013	0.026	-0.029	0.78	-0.027	0.22	-0.0059	0.018	1	-0.022	0.015	0.63	0.46	0.41	0.46	-0.068
TrainingTimesLastYear	-0.016	-0.032	-0.014	0.0012	-0.024	-0.025	-0.00039	-0.028	-0.051	-0.013	0.026	-0.022	1	0.037	0.0086	0.01	0.019	-0.013	-0.0054
EmpWorkLifeBalance	-0.02	-0.045	0.01	-0.00026	0.016	-0.014	0.049	-0.019	0.0025	-0.017	-0.0049	0.015	0.037	1	0.023	0.046	0.016	0.011	0.12
ExperienceYearsAtThisCompany	0.32	0.022	0.076	-0.00056	-0.0004	-0.04	0.54	0.0018	-0.13	-0.02	0.016	0.63	0.0086	0.023	1	0.76	0.62	0.76	-0.11
ExperienceYearsInCurrentRole	0.22	0.02	0.067	0.025	-0.012	0.0029	0.4	0.002	-0.097	-0.005	-0.032	0.46	0.01	0.046	0.76	1	0.54	0.73	-0.15
YearsSinceLastPromotion	0.23	0.013	0.054	0.011	-0.01	-0.02	0.36	-0.0065	-0.032	-0.016	0.026	0.41	0.019	0.016	0.62	0.54	1	0.49	-0.17
YearsWithCurrManager	0.21	0.018	0.089	-0.012	-0.0046	0.013	0.37	-0.022	-0.11	-0.0077	-0.0064	0.46	-0.013	0.011	0.76	0.73	0.49	1	-0.12
PerformanceRating	-0.04	-0.046	0.021	0.4	-0.043	-0.011	-0.077	0.00061	0.021	0.33	-0.02	-0.068	-0.0054	0.12	-0.11	-0.15	-0.17	-0.12	1

We can see that these variables play an important role that affects the performance rating of the employee (They have a strong co-relation).

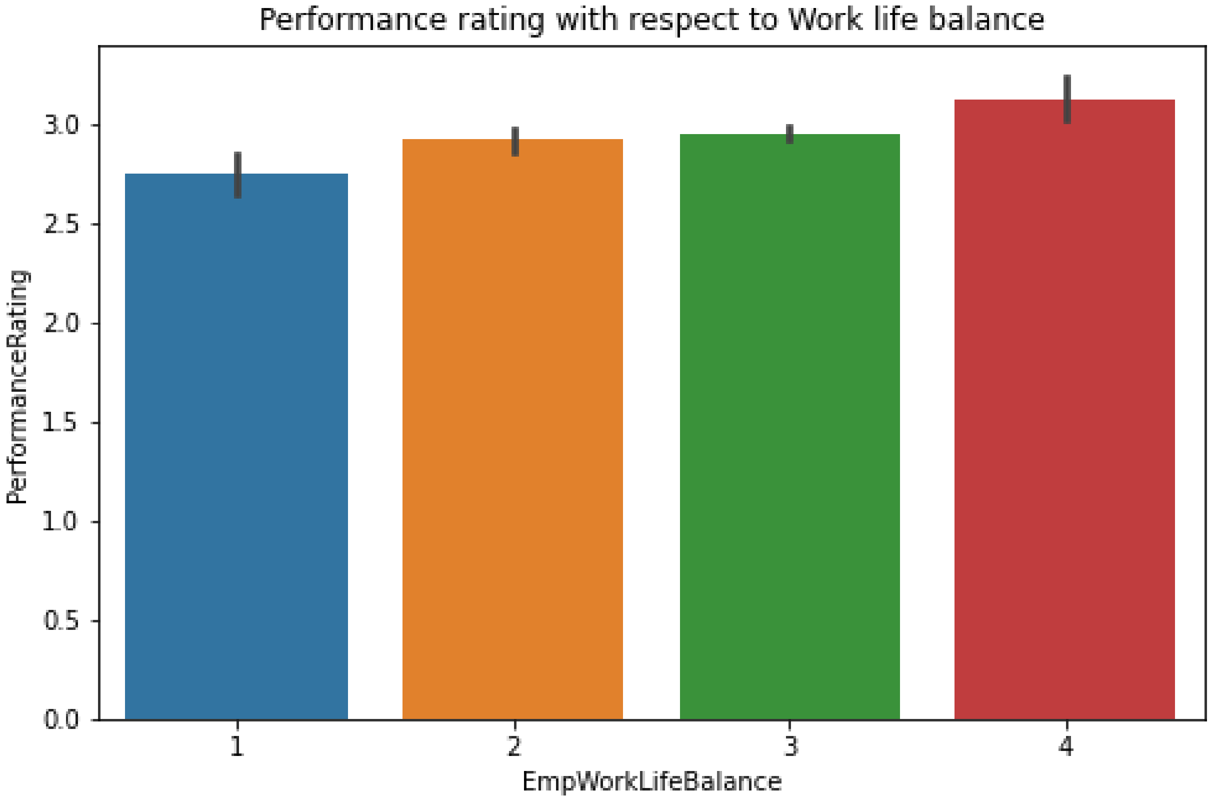
- Employment Environment Satisfaction
- Work life balance
- Last Salary Hike Percent & Work life balance

Let us perform an analysis on these variables and understand more about the factors that affect's the employee's performance rating.

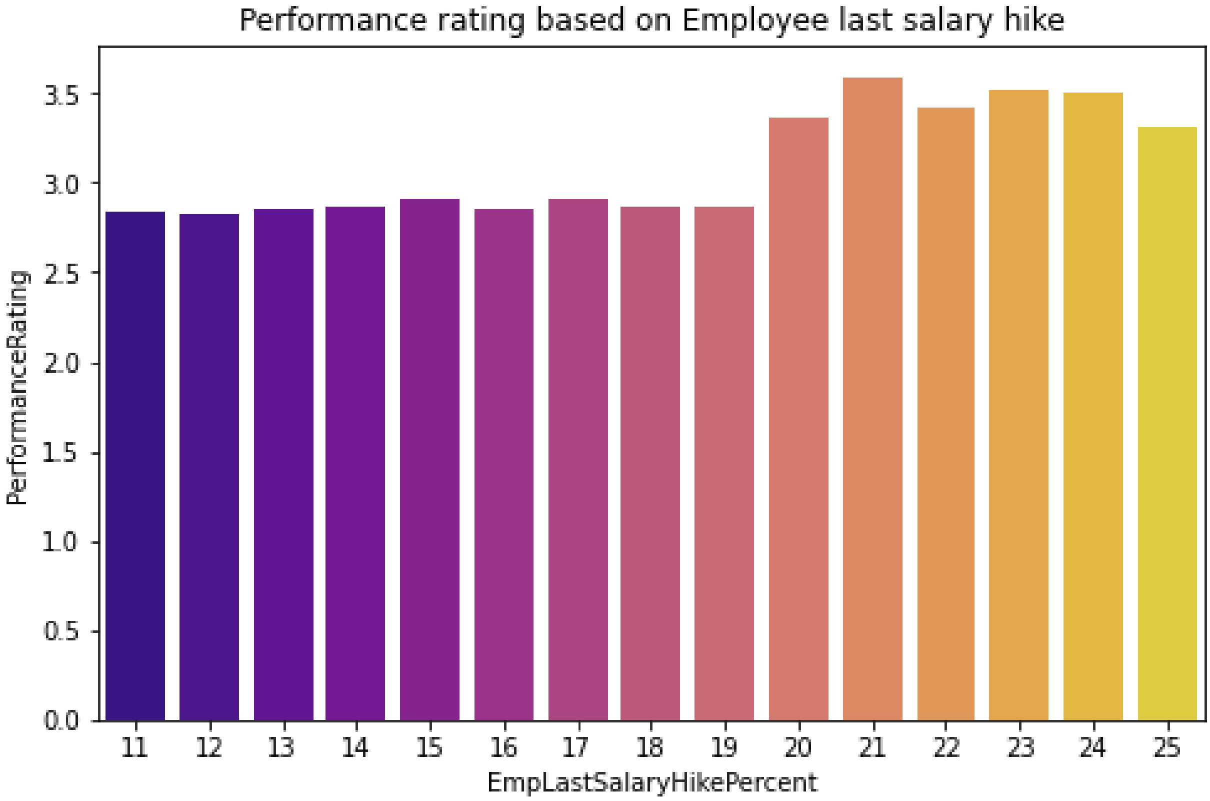
Performance Rating based on Environment satisfaction



Performance based on worklife balance



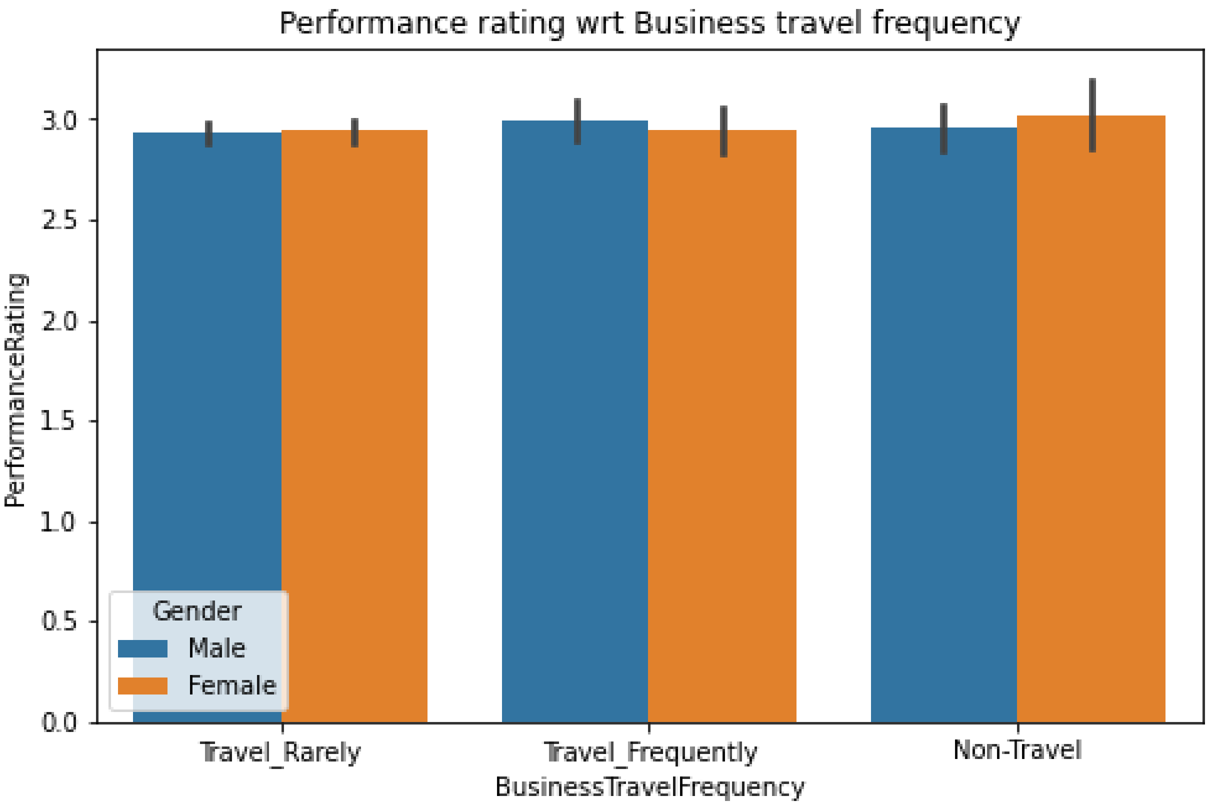
Performance based on Last salary hike percent

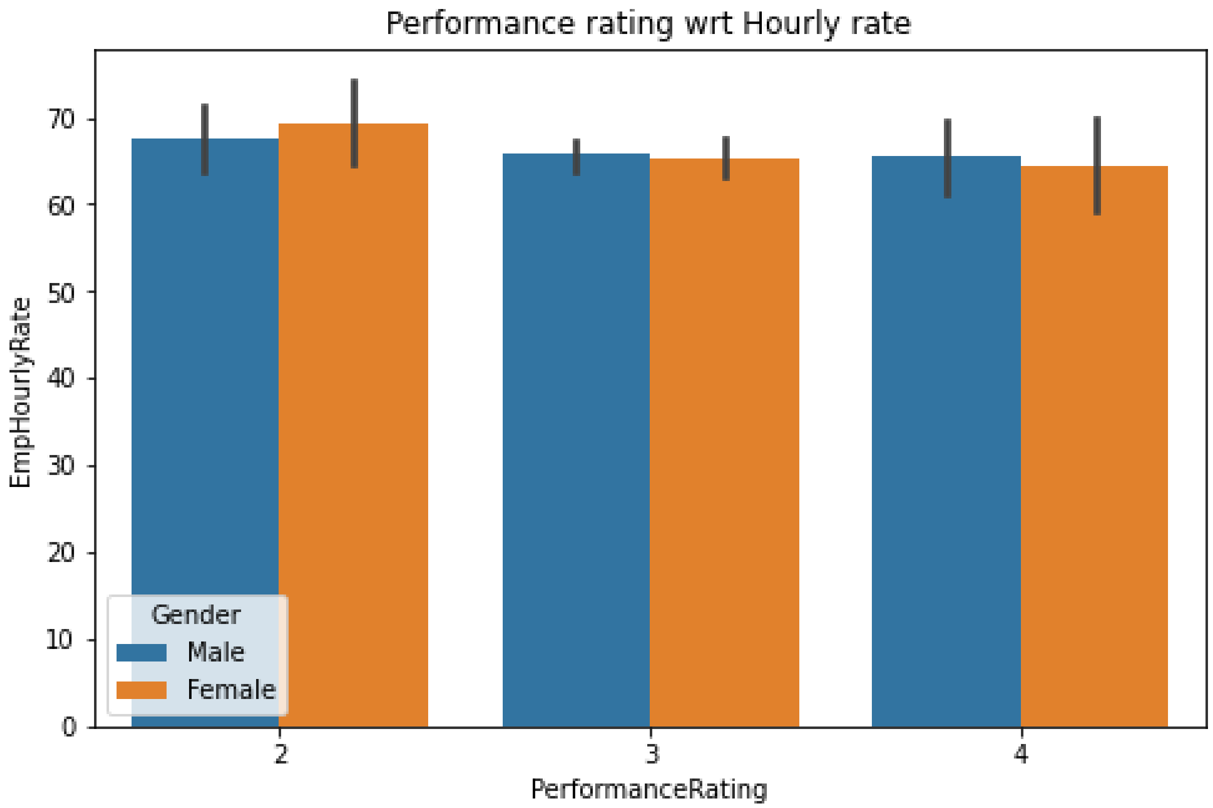


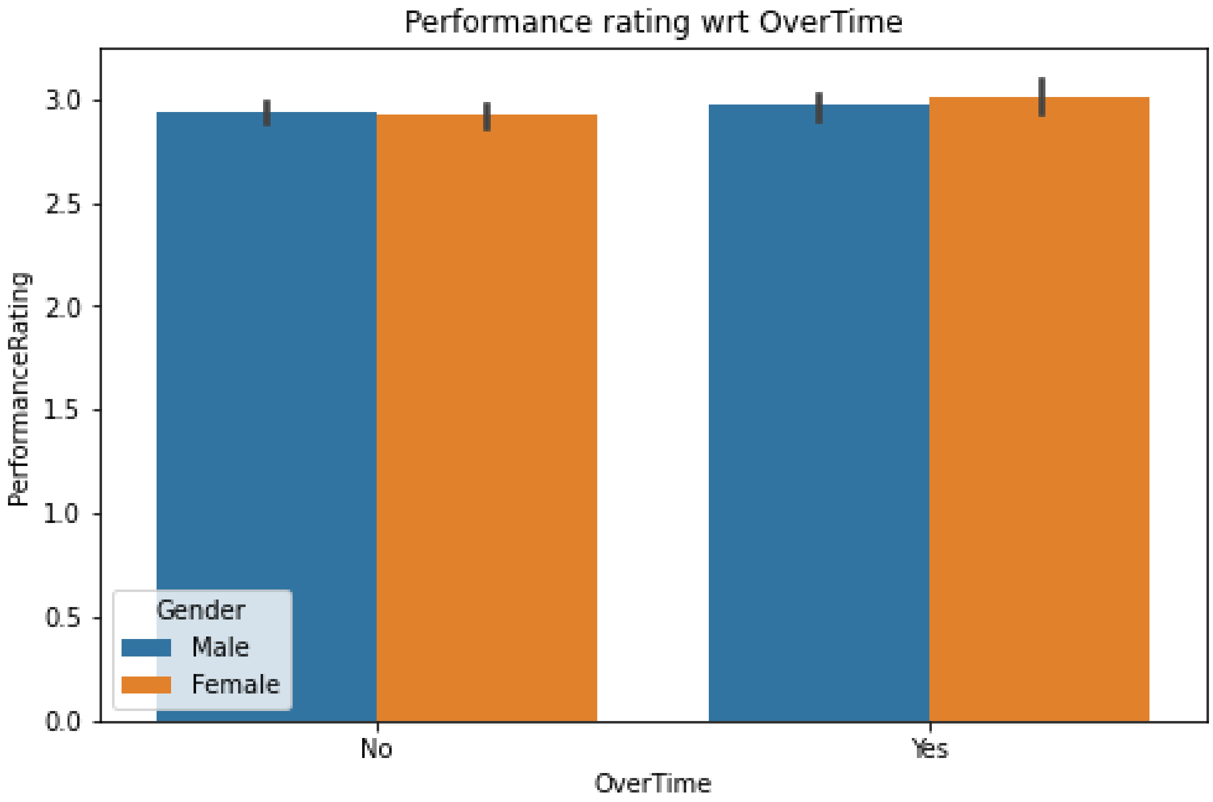
A salary hike can also motivate the employee's to perform well , we can see that 20-25 have a decent performance rating

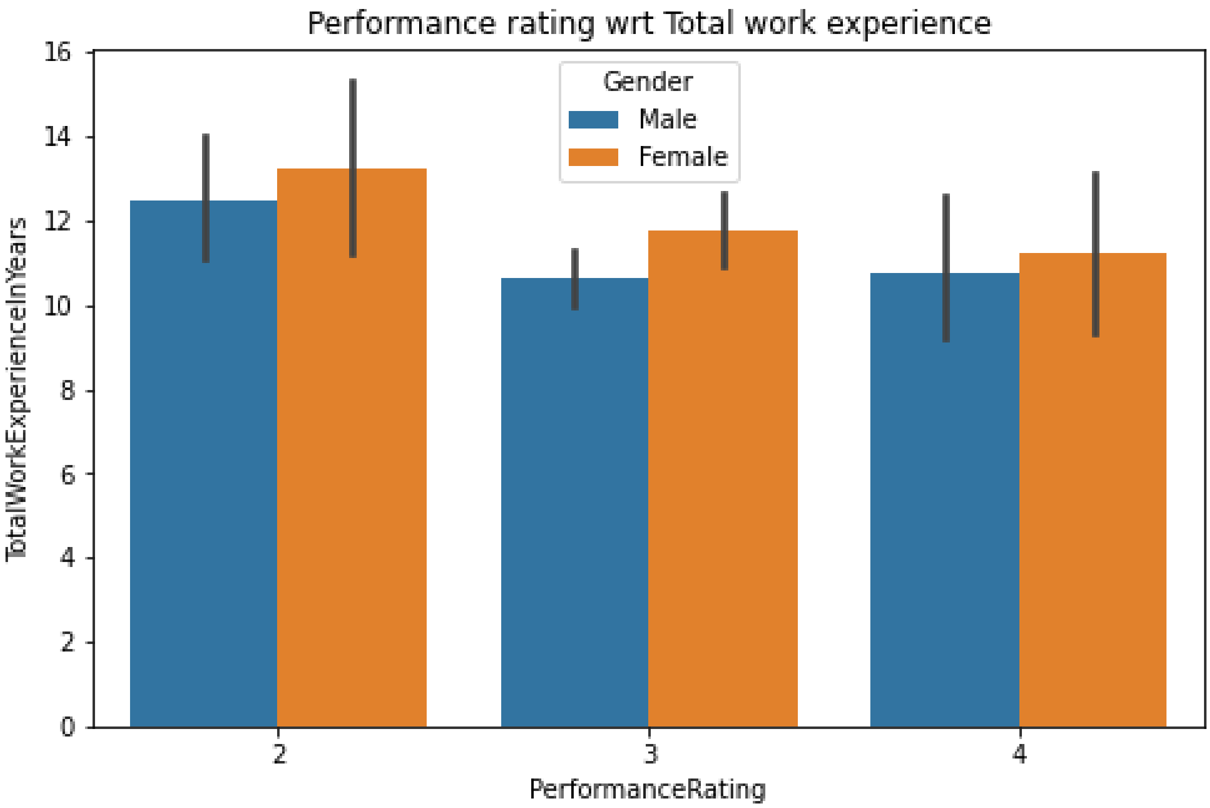
Other factors

Now we will look at other factors that might affect the performance rating of the employee.

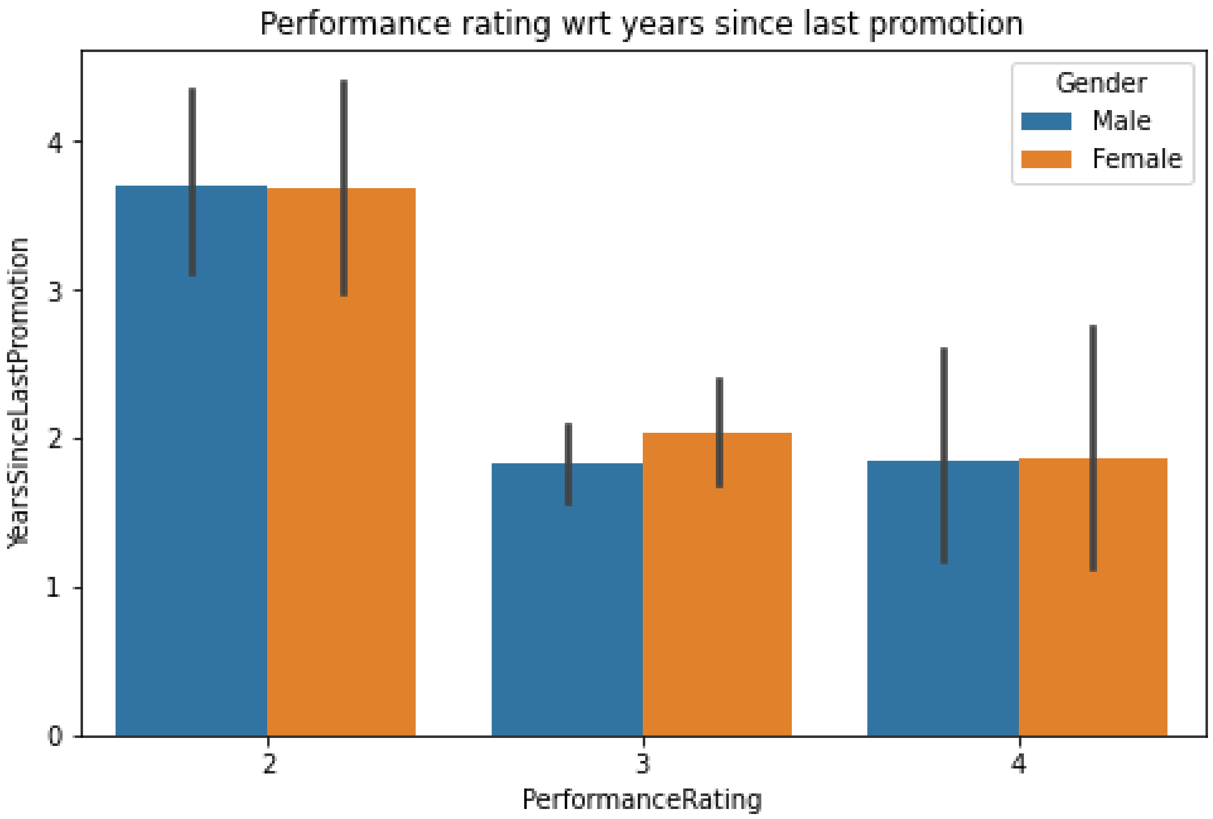

Performance rating wrt Business travel frequency
PerformanceRating
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Gender
Male
Female
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravelFrequency

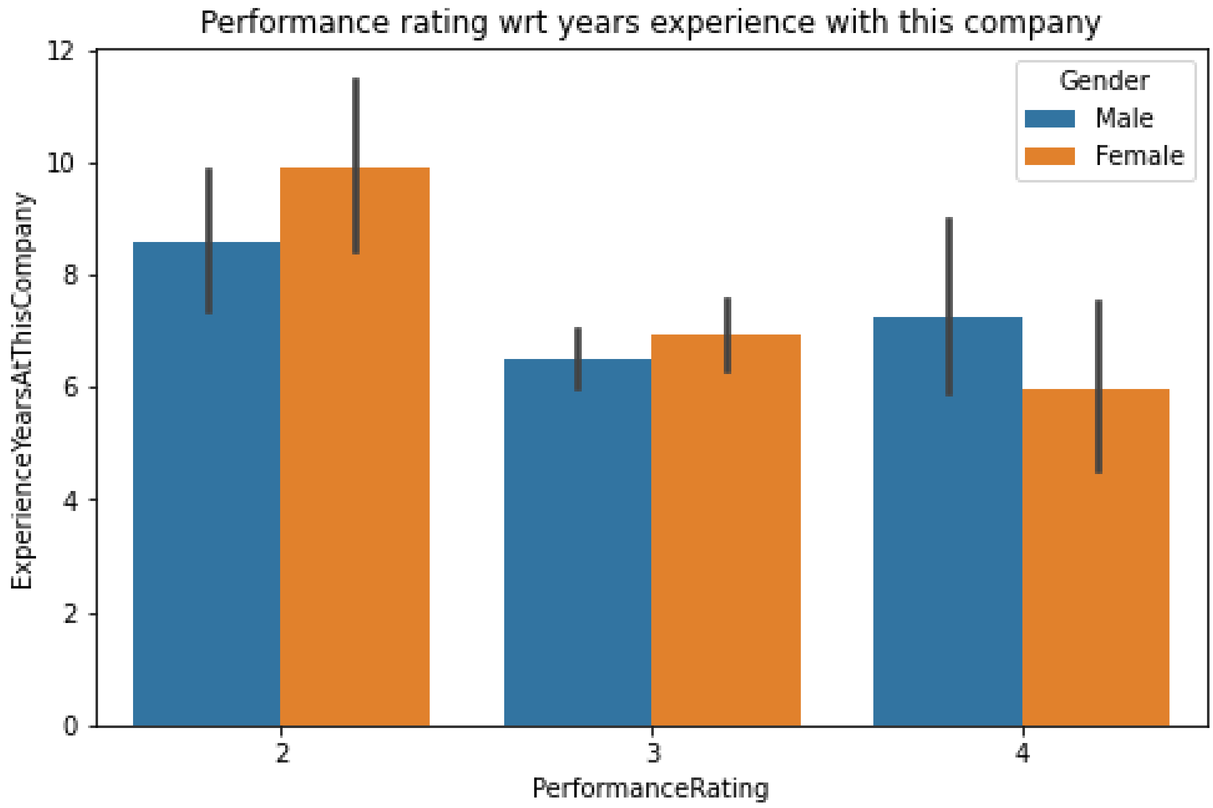

Performance rating wrt Hourly rate
EmpHourlyRate
0
10
20
30
40
50
60
70
PerformanceRating
2
3
4
Gender
🟦 Male
🟧 Female


Performance rating wrt OverTime
PerformanceRating
0.0
0.5
1.0
1.5
2.0
2.5
3.0
OverTime
No
Yes
Gender
🟦 Male
🟧 Female

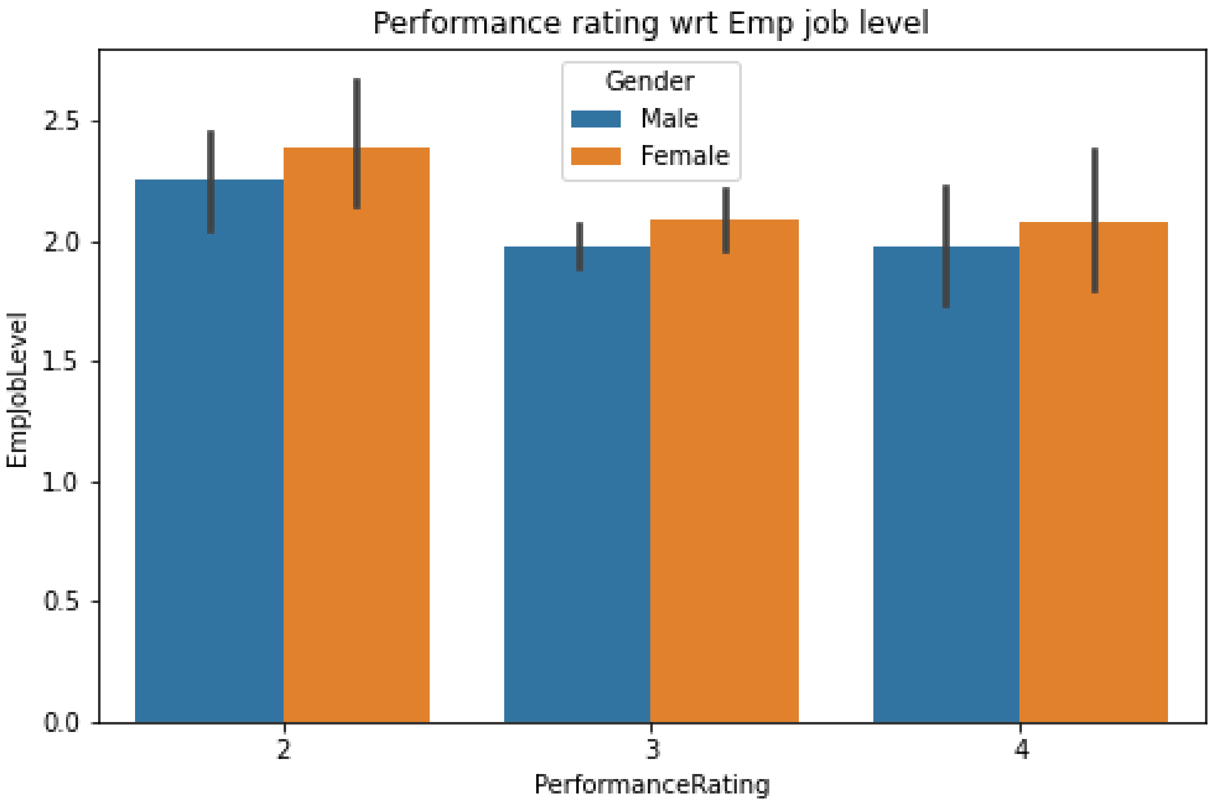


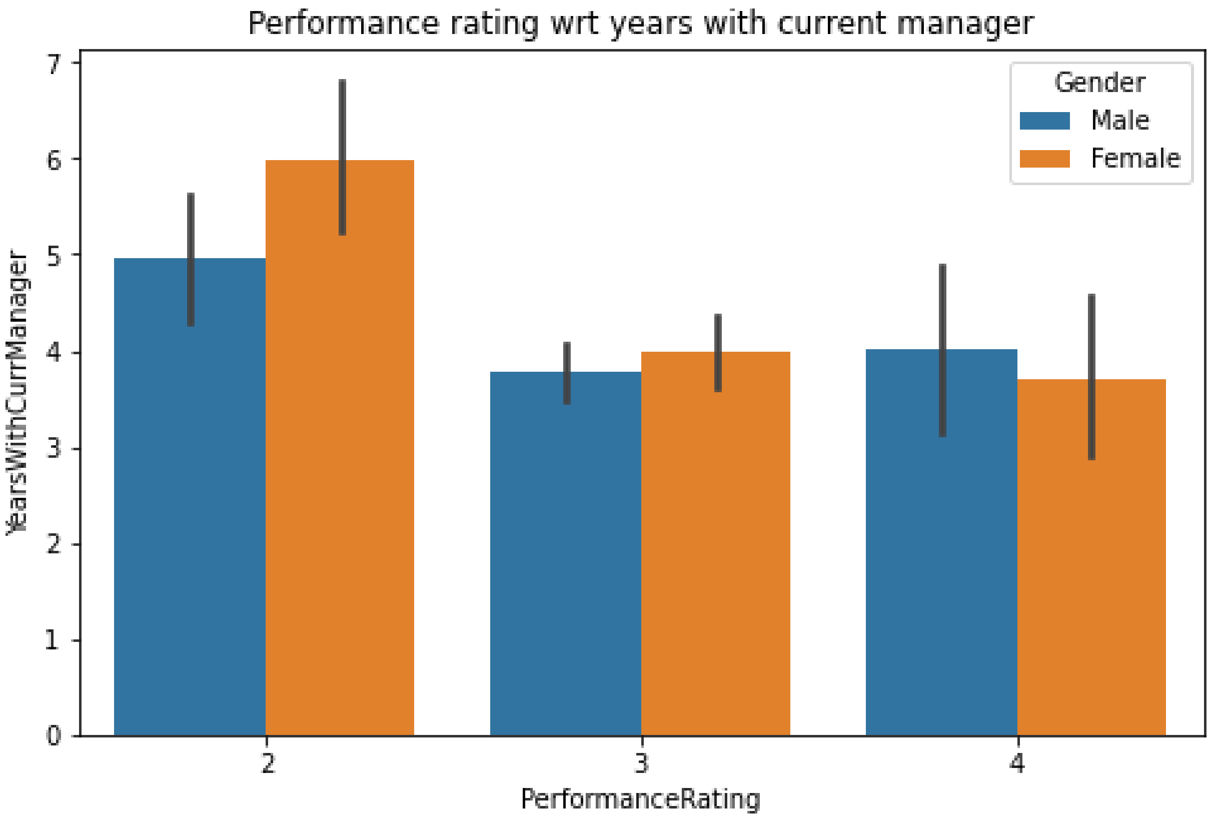
We can see that people with higher experience have low performance rating , and people with lower year's of experience have good performance rating





Higher the experience lower the performance rating , probably because of the pressure that the higher experience of employee has & the work life balance


Performance rating wrt Emp job level
Gender
Male
Female
EmpJobLevel
0.0
0.5
1.0
1.5
2.0
2.5
2
3
4
PerformanceRating


Performance rating wrt years with current manager
Gender
Male
Female
YearsWithCurrManager
0
1
2
3
4
5
6
7
2
3
4
PerformanceRating

Results

- Using Feature Engineering in Random Forest Classifier,

Accuracy Score was 92.91%, Precision Score was 92.71% and Recall Score was 92.91%.

- Using Randomized Search Cross Validation in Random Forest Classifier,

Accuracy Score was 80%, Precision Score was 83.33% and Recall Score was 80%.

- **Feature Engineering** technique results in highest accuracy score, precision score and recall score.
- **Randomized Search Cross Validation** techniques result in lowest accuracy score, precision score and recall score.
- Also **Random Forest Classifier**, results in more than 90% accuracy, precision and recall.

Recommendations to improve the employee performance

- I recommend that the company should focus on the three factors that affect employee performance i.e. Employee Environment Satisfaction, Last Salary Hike Percent and Years Since Last Promotion and improve on them.
- It means that employee needs to be happy on the job.
- The salary of the employees needs to be raised twice a year and those who perform better needs to be promoted every year. This will in turn boost the confidence of the employees.
- The other factors like Experience Years in Current Role (14.76%), Employee Work Life Balance (12.44%), and Years with Current Manager (12.23%) also need to be carefully monitored for better functioning of the organisation.
- Males need to work hard in order to be in par with Females with respect to Performance.
- They need to improve in Human Resources and Finance Departments.
- Females need to improve in Sales and Finance Departments.
- Finance Department needs to closely monitor their employees as both males and females have not performed better.

1. The algorithm and training method(s) you used (Such as SVM, Neural Network etc.,)

Ans) The algorithms used for this project are

1. Random Forest Classifier

2. The most important features selected for analysis (Whether techniques such as PCA Factorization used)

The most important features selected for analysis were

1) EmpDepartment, 2) Gender, 3) BusinessTravelFrequency, 4) DistanceFromHome, 5) EmpEducationLevel, 6) EmpEnvironmentSatisfaction, 7) EmpJobInvolvement, 8) EmpJobLevel, 9) EmpJobSatisfaction, 10) OverTime, 11) EmpLastSalaryHikePercent, 12) NumCompaniesWorked, 13) EmpRelationshipSatisfaction, 14. TrainingTimesLastYear, 15) EmpWorkLifeBalance, 16) ExperienceYearsAtThisCompany, 17) ExperienceYearsInCurrentRole, 18) YearsSinceLastPromotion, 19) YearsWithCurrManager, 20) EmpHourlyRate, 21) Attrition

All these features were selected because they form the major part in employee appraisal as well as company's performance.

The techniques used in the analysis were

1. Randomized Search Cross Validation (CV) in Random Forest Classifier: Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. The chances of finding the optimal parameters are higher in random search because of the random search pattern where the model ends up being trained on the optimised parameters.

2. Label Encoder for Data Processing and Data Munging: To convert the categorical data into numerical data so that it is easier for predictive models to understand the data.

3. Other techniques and tools used in the project.

The other techniques used in this project are

1. Scaling Technique: It standardizes the dataset on any axis.
2. Standard Scaling Technique: It standardizes features by removing the mean and scaling to unit variance.
3. Feature Engineering Technique: It uses the domain knowledge to extract features from the raw data. These features help to improve the performance of the machine learning algorithms.

Tools

Packages used in this project are **matplotlib**, **pyplot**, **seaborn**, **numpy**, **pandas**, **sklearn**, **collections**, **imblearn**,

1) **important feature transformations**

outliers were present in some of the fields like

1. TotalWorkExperienceInYears
2. ExperienceYearsAtThisCompany
3. YearsSinceLastPromotion

They were removed and new features were created (after transformation) as follows:

1. clean_TotalWorkExperienceInYears
2. clean_ExperienceYearsAtThisCompany
3. clean_YearsSinceLastPromotion

Then the features which were present before transformation were dropped from the dataframe. This technique in turn results in much better accuracy for all the algorithms that are used in this project.

Correlation or interactions among the features selected

correlation was selected among the features using python code **corr = data.corr()** where performance is the

dataframe containing the data of employee performance prediction.

This will help used to find the pairwise correlation of all columns in the dataframe. It generates a correlation matrix.

Also we have to use heatmap to get the visual representation of correlation matrix which is supported by package called seaborn.

interesting relationships in the data that don't fit in the sections above

The fields like ExperienceYearsAtThisCompany, ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager are most positively correlated against each other compared all the other fields. It is depicted in yellow color rectangle as in Correlation heat map.

most important technique you used in this project

The most important technique used in this project is Feature Engineering in Random Forest Classifier by sorting the values based on correlation with Performance Rating. The results were 1. Accuracy = 92.91%, 2. Precision score = 92.71%, 3. Recall score = 92.91%.