Name:	Susan Mumbi Kisengeu
Email:	mkisengeu@gmail.com
Assessment ID:	E10901-PR2-V18
Module:	Certified Data Scientist - Project
Submission Date:	14th April 2024

Table of Contents

1. Introduction

INX Future Inc. is one of the leading data analytics and automation solutions provider with over 15 years of global business presence. Over the past 5 years, the company has emerged as top 20 best employers with employee-friendly human resource policies. However, a recent survey showed an 8% decrease in service delivery and client satisfaction levels. Through the Chief Executive Officer (CEO), Mr. Brain, this data science project was thus initiated to analyze the current employee data and find the underlying causes of the performance issues. The findings of this project will be used to implement the right cause of action, ensuring that the non-performing employees are penalized without affecting the overall employee morale. Therefore, the undertakings of this project aims to give an overall insight on the factors affecting the performance rating of employees, and create a prediction model for future hires using machine learning.

2. Objectives

The goals and the insights for this project are:

i. To analyze department wise performances.
ii. To identify the top 3 important factors effecting employee performance.
iii. To create a trained model which can predict the employee performance based on factors as inputs, for use in hiring employees.
iv. To give recommendations for improving the employee performance based on insights from analysis.

3. Data Collection and Preparation

3.1 Data Source

The project script is done on Jupyter Notebook (.ipynb format) and the analysis is based on data from a third party, accessible via the link:
http://data.iabac.org/exam/p2/data/INX_Future_Inc_Employee_Performance_CDS_Project2_Data_V1.8.xls

The dataset composition is as follows:

- Number of employee entries: 1200 (rows)
- Number of features: 28 (columns)

The dataset has a combination of 19 Numerical and 9 Categorical Features as shown in *Table 1* below. The numerical features comprise of discrete, continuous or time-series based values; while categorical features contain nominal, ordinal, ratios or interval-based values. For instance, the employee number data is alphanumerical and serves as an identifier (ID), not affecting the relevant features such as the performance rating.

Table 1: Features in the Dataset

Numerical Data	Categorical Data
1. Age 2. Distance from Home 3. Employee Education Level 4. Employee Environment Satisfaction 5. Employee Hourly Rate 6. Employee Job Involvement 7. Employee Job Level 8. Employee Job Satisfaction 9. Number of Companies Worked 10. Employee Last Salary Hike Percentage 11. Employee Relationship Satisfaction 12. Total Work Experience in Years 13. Training Times Last Year 14. Employee Work-Life Balance 15. Years of Experience at this Company 16. Years of Experience in Current Role 17. Years since Last Promotion 18. Years with Current Manager 19. Performance Rating	1. Employee Number 2. Gender 3. Education Background 4. Marital Status 5. Employee Department 6. Employee Job Role 7. Business Travel Frequency 8. Overtime 9. Attrition

3.2 Analysis Tools

- Python: Utilized for data analysis, machine learning modeling, and visualization.
- Libraries used: pandas, NumPy, scikit-learn (sklearn), Matplotlib, Seaborn and Streamlit.
- Machine Learning Models and Techniques considered in this analysis: Random Forest Classifier, Decision Trees; Feature Selection using ANOVA F-value.
- Data Analytics and Evaluation Tools: Exploratory Data Analysis (EDA), Cross-Validation, Confusion Matrix, and Classification Metrics (Precision, recall, F1-score, and accuracy).

4. Exploratory Data Analysis

4.1 Univariate Analysis

To get an overall view of the dataset, univariate analysis was used to explore each variable, separately. It led to the following general observations for the **numerical features** (*see sample graphical illustrations in Appendix 1a)*:

a. The age of the employees in the dataset ranges from 18 – 60, with a majority being between 30-40 years.
b. A majority of employees in the dataset have short commute distances to work (1-12 km).
c. More than half the employees are satisfied with their work environment giving a rating above 3.0.
d. Also, slightly more than half the employees in the dataset have worked in 0-1 companies, and a majority of the employees have worked for between 0-10 years in INX.
e. About 3-quarters of the employees in the dataset have a performance rating of 3, with the other quarter having a rating of 2 or 4.

Similarly, for the **categorical features**, the following observations were made with regard to the dataset (*see sample graphical illustrations in Appendix 1b)*:

a. The number of males (>700) is higher than the number of females (<500).
b. A majority of employees rarely travel.
c. About 1/3 of the employees have overtime.
d. A majority of the employees are Sales Executives, followed by Developers.
e. A majority of the employees are married.
f. More than half the company have an educational background in Life Sciences and Medical.

These overall observations give an insight on the type of company and information being assessed. However, in order to get answers to the objectives/ goals of the project, further analysis is required, such that two relatable features can be compared as shown in section 4.2 below.

4.2 Bivariate Analysis:

This is a way of using statistical/ descriptive methods to give a relationship between components X and y, for instance. Some common methods include: *analysis of variance (ANOVA) test, sample t-test, Scatterplots, Correlation, etc.* [1].

To address the ***1st objective*** of the study, a simple descriptive statistics of the dataset is used to visualize the interaction of the performance rating and the departments. This is done using the function $df.describe()$ for numerical averages and ***Bar-plots*** for visualization. *Table 2* below shows a sample of the descriptive statistics for the first 3 columns (age, distance from home, and education level). This gives the mean (averages), standard deviation, maximum and minimum values, etc.

Table 2: Descriptive Statistics of the Dataset

	Age	DistanceFromHome	EmpEducationLevel
count	1200.000000	1200.000000	1200.00000
mean	36.918333	9.165833	2.89250
std	9.087289	8.176636	1.04412
min	18.000000	1.000000	1.00000
25%	30.000000	2.000000	2.00000
50%	36.000000	7.000000	3.00000
75%	43.000000	14.000000	4.00000
max	60.000000	29.000000	5.00000

The statistical approach addresses ***Objective 1***, "To analyze department wise performances", by giving an overall average performance rating per department as shown in the graphical illustration shown in Figure 1 below.

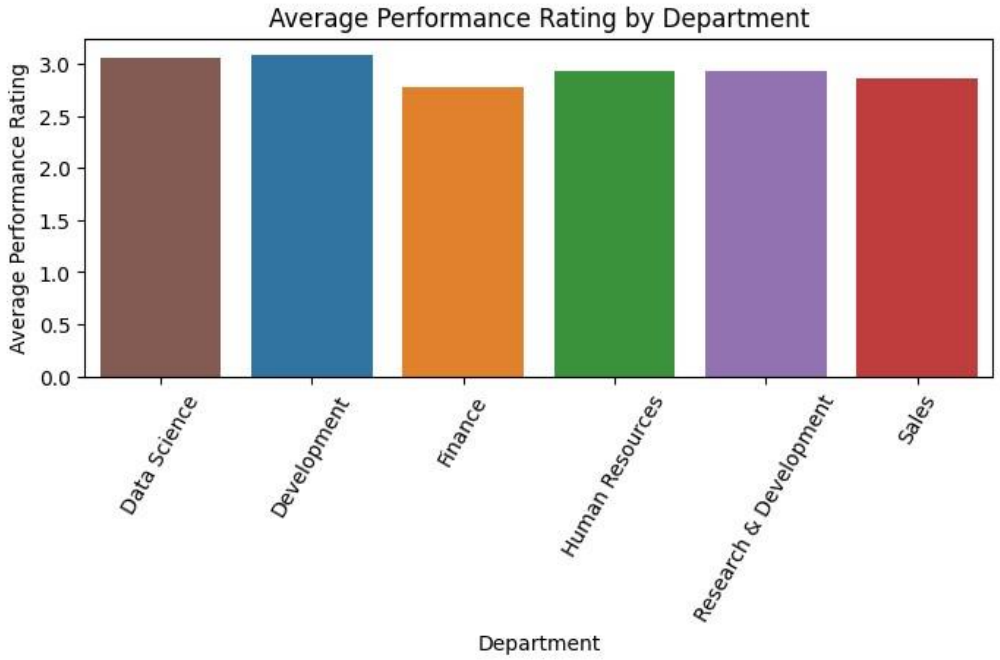


Figure 1: Average Performance Rating Per Department

The departments can thus be ranked as per their numerical averages as shown in Table 3 below:

Table 3: Department-wise Performance Rating (Numerical Averages)

Dept. Index	Employee Department	Performance Rating
1	Development	3.0859
0	Data Science	3.0500
3	Human Resources	2.9259
4	Research & Development	2.9213
5	Sales	2.8606
2	Finance	2.7755

From the above averages, the best rated department is "*Development*", closely followed by "*Data Science*", and the least performing department is "*Finance*". In addition to the averages, the count of employees in each department is represented in Figure 2 below, showing that over 300 employees in the "Development" department were rated 3, followed by "Sales" and "Research & Development". The least performing employees were rated 2, that is over 80 in "Sales", over 60 in "Research & Development", and less than 10 in "Finance", "Human Resource" and "Data Science". The statistics gives an overall performance rating per department, considering that all departments have a different job roles and employee count.

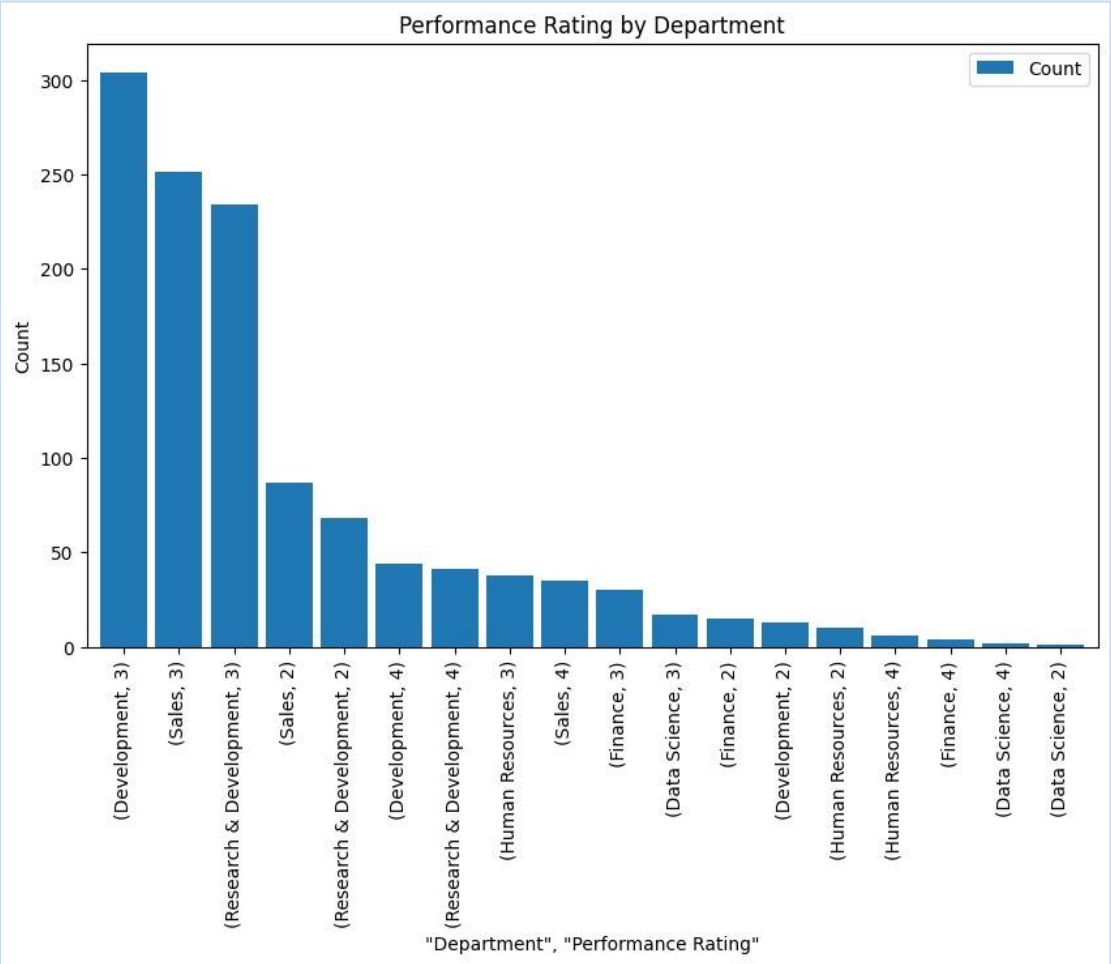


Figure 2: Performance Rating per department

4.3 Correlation Analysis:

For **Objective 2**, "*To identify the Top 3 Important Factors effecting employee performance*", the bar-plots and descriptive statistics need to be comparable for all features of the dataset. The analysis can be done either through a machine learning model (i.e. following steps shown in Sections 4.2 and 4.3 below), or from an initial generic point of view (through a correlation matrix or scatter plots). In this project, the correlation matrix was obtained for all 28 features of the dataset to first show a holistic

view of the interrelations of the features. A correlation matrix gives positive, negative or zero value to imply:

- Positive – linear relation between two variables.
- Negative – inverse relation between two variables.
- Zero – no relation between two variables [2].

Figure 3 below shows the overall correlation matrix for the dataset, considering *all features*:

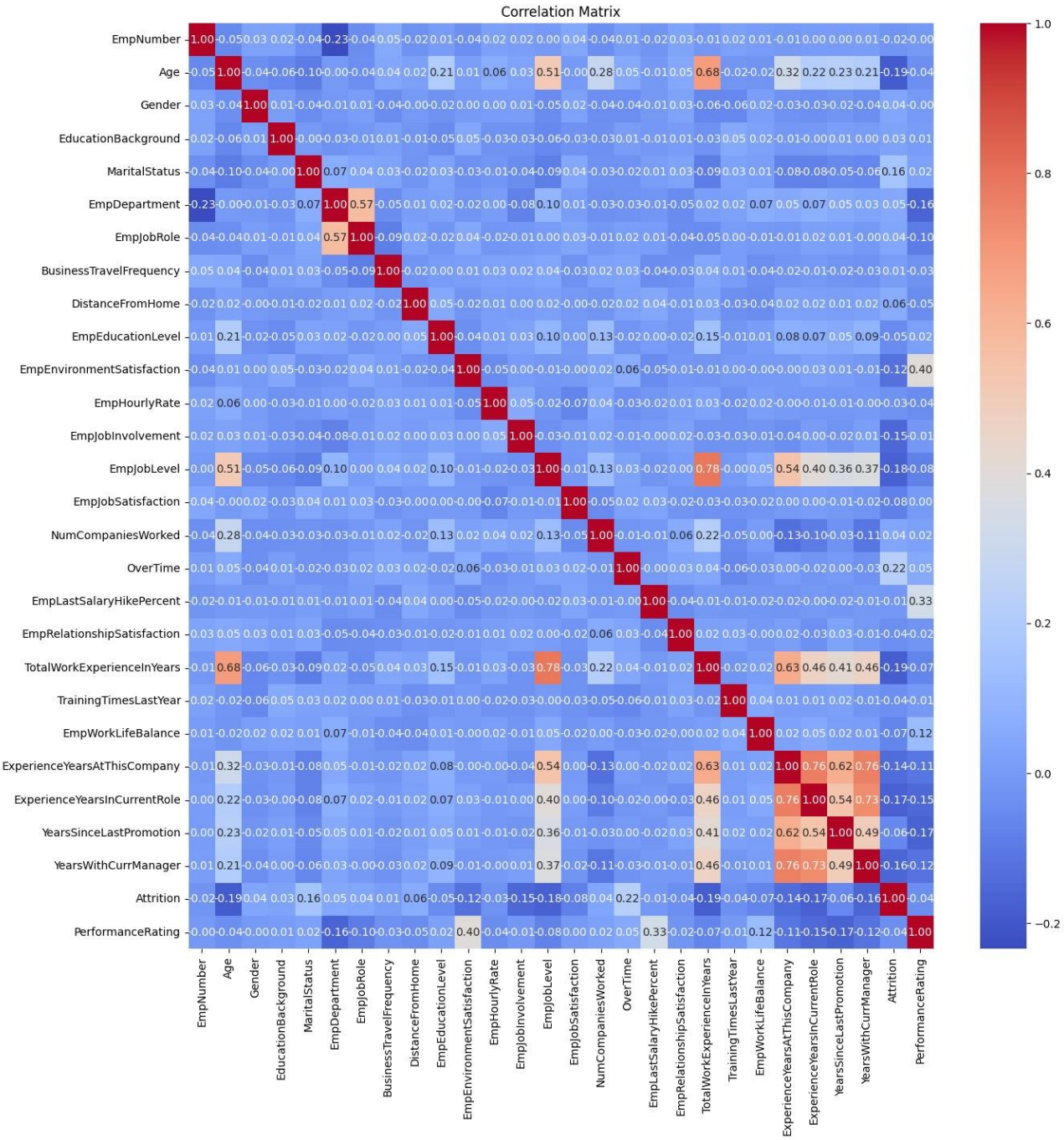


Figure 3: Correlation matrix for all features

In this project, a heat-map correlation was used since the matrix can get complex when using a lot of features. The preference over the scatterplots was in the obtaining of a unifying graphical illustration. From the correlation matrix above:

- The center **red** diagonal with a value of 1.00 represents the relation of each feature with itself, which is normally ignored.
- The dark **blue** shade ranging from 0 to - 0.20 shows inversely related such as "age with attrition rate", "employee department with performance rating".
- High positive values show a direct relation, e.g., **0.78** shows a direct relation between the "total years of work experience with the employee's job level"; or **0.68** for an employee's age being directly related to the number of years of experience. These are both practical conclusions even in the real-world.

Therefore, to address the objective of the top 3 factors affecting employee performance, the values of the relation between the performance rating with the other features is ranked in *Table 3* below:

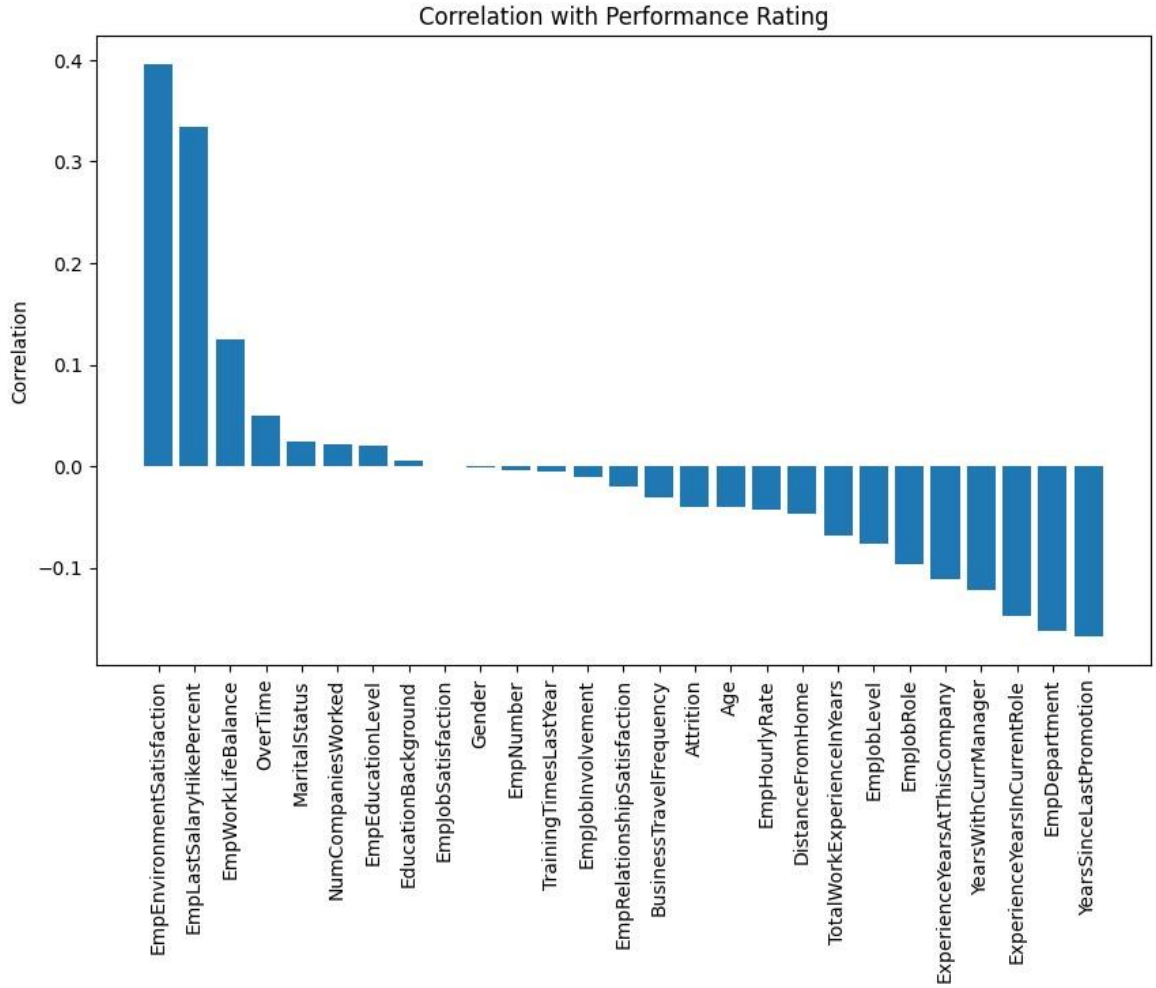
Table 4: Correlation of Features with Performance Rating.

#	Features	Correlation Values
1	Employee Environmental Satisfaction	0.3956
2	Employee Last Salary Hike Percentage	0.3337
3	Employee Work-Life Balance	0.1244
4	Over-Time	0.0502
5	Marital Status	0.0242
6	Number of Companies Worked	0.0210
7	Employee Education Level	0.0205
8	Education Background	0.0056
9	Employee Job Satisfaction	0.0006
10	Gender	-0.0018
11	Employee Number	-0.0032
12	Training Times Last Year	-0.0054
13	Employee Job Involvement	-0.0105
14	Employee Relationship Satisfaction	-0.0195
15	Business Travel Frequency	-0.0310
16	Attrition	-0.0398
17	Age	-0.0402
18	Employee Hourly Rate	-0.0431
19	Distance From Home	-0.0461
20	Total Work Experience In Years	-0.0681
21	Employee Job Level	-0.0766
22	Employee Job Role	-0.0962
23	Experience Years At This Company	-0.1116
24	Years With Current Manager	-0.1223
25	Experience Years In Current Role	-0.1476
26	Employee Department	-0.1626
27	Years Since Last Promotion	-0.1676

The ***top 3 positively correlated*** features are the employees' *environmental satisfaction, last salary hike percentage,* and *work-life balance*. However, on the ***negative correlation***, the employees' *years of experience in the current role*, *department,* and *years since last promotion* indicate an inverse relation with the Performance rating. Therefore, taking the absolute values, features 1, 2 and 27 would affect performance the most. That is:

- The ***more*** the satisfaction with the work environment, the ***higher*** the performance rating.
- The ***higher*** the last salary hike percentage, the ***higher*** the performance rating.
- The ***more*** the years since the last promotion, the ***less*** the performance rating.

The above results can also be represented in graphical form as shown below:



A high negative correlation could also indicate areas of concern / potential areas contributing to lower employee performance. These can be recommended as areas of improvement within INX organization as highlighted in the Recommendations in Section 5.

4.4 Features Engineering

Data Pre-Processing

For specific data analysis, data can be preprocessed so as to detect and correct (or remove) corrupt or inaccurate records from a dataset. In the initial analysis of the dataset above, the following was checked:

i. Missing values – no null/ NaN values in the dataset.
ii. Duplicated values – none.

Since most Machine Learning (ML) models accept numerical variables only, preprocessing the categorical variables is also required. The categorical data in this dataset is ordinal (i.e., categorical variables with a natural order), therefore, ***Label Encoding*** is used to assign numerical labels based on the order or ranking of the feature, e.g. Male vs Female in Gender. The project uses the class from the $sklearn.preprocessing$ module for this process, converting the 9 categorical features [3]. Checking the data frame info using the $df.info()$ command, the data is confirmed to be of numerical / integers datatype, thus ready for ML modeling [4].

Feature Selection

This is the process of removing features that are not useful for classification. These features can be removed based on the following criteria:

- *Correlation / Redundancy* - e.g. Employee Job Role and Department are correlated (*correlation value 0.57*) and one of them can be dropped.
- *Business importance* - from the above example, the department can be dropped and in its place, the job role can be used.
- *Feature importance* - can be analyzed using the ANOVA F-value test to rank the most important features.

Feature selection thus reduces the input variables to the model by using only relevant data and getting rid of noise in data. The *Selection of the "Best" Features* using the $Select - K - Best$ technique in $sklearn$ was used, which utilizes the ANOVA F-value to pick the top 10 values. The scoring function evaluates how well each feature relates to the target variable (performance rating). Generally, the higher the score for a feature, the more relevant it is considered[4], [6].

From the selection criteria above, the top 10 features are:

1. Employee Department
2. Environmental Satisfaction
3. Overtime
4. Last Salary Hike Percentage
5. Work Life Balance
6. Number of years of experience at the company
7. Number of years of experience in the current role
8. Years since last promotion
9. Years with current manager
10. Performance Rating

However, in the model chosen in section 4.4 below, some of the features like Overtime, doesn't come in the feature importance ranking shown in *Figure 4* below:

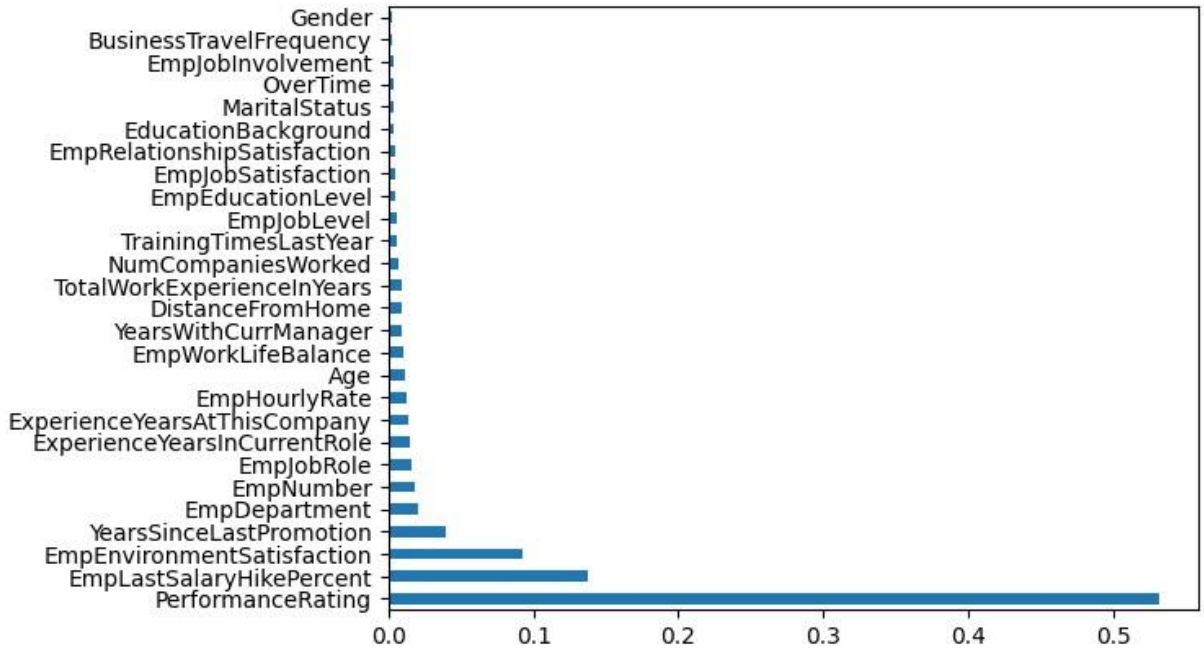


Figure 4: Feature Importance Ranking

Therefore, the feature selection can vary based on the model chosen. In this project, we compared the performance of using both both selected features and all features, which showed a minimal deviation for solution accuracy as shown in section 4.4 below.

4.5 Predictive Modeling

So far, the process being followed in modeling the machine learning model is as per the following steps [7]:

i. **Define the problem:** To create a trained model which can predict the employee performance based on factors as inputs to be used to hire employees **(Objective 3)**.
ii. **Collect the data:** INX Future Inc. excel file.
iii. **Prepare the data:** Feature Engineering, split data into train and test sets (30%), etc.
iv. **Explore the data**: use previous exploratory data analysis outputs to explain relationships between variables.
v. **Model the data:** e.g. using the Random Forest Classifier due to its high classification accuracy.
vi. **Evaluate the model:** use the test set of data to evaluate the model. Then check the accuracy of the model, check for overfitting, precision, F1 score, etc.
vii. **Deploy the model**: the model can then be used to make predictions on new hires (Streamlit app).

The models factored in for use in this project include:

i. Decision Tree
ii. Random Forest Classifier.
iii. Gradient Boosting Classifier
iv. Logistic Regression (for multi-classification).

The **Random Forest Classifier** was chosen based on its simple implementation. Classifications models are considered in place of Regression models since the output required (Performance Rating) is a discrete value (2,3,4), whereas for the regression models, the output would be a continuous value [8]. The parameters below were considered:

- Number of estimators=100,
- Maximum depth – comparison of 5 and 10
- Minimum sample split = 2 and 3
- Random state set to 123

Achieving an output such as Figure 5 below:

```
test accuracy:  0.9861111111111112
all features test accuracy:  0.9972222222222222
              precision    recall  f1-score   support

           2       1.00      1.00      1.00        49
           3       0.98      1.00      0.99       268
           4       1.00      0.88      0.94        43

    accuracy                           0.99       360
   macro avg       0.99      0.96      0.98       360
weighted avg       0.99      0.99      0.99       360

              precision    recall  f1-score   support

           2       1.00      1.00      1.00        49
           3       1.00      1.00      1.00       268
           4       1.00      0.98      0.99        43

    accuracy                           1.00       360
   macro avg       1.00      0.99      1.00       360
weighted avg       1.00      1.00      1.00       360
```

Figure 5: Random Forest Classifier Sample Output

Model Output:

- From *Figure 5* above, *all features* had a higher accuracy of 99.72% as compared to the *selected features* (accuracy of 98.61%). A weighted average of 99% for selected features and 100% for all features shows that the latter achieved higher prediction and test accuracy across all classes.
- In additional tests, the model achieved high test accuracy, indicating its effectiveness in predicting the target variable for the instances in the test dataset. With a maximum tree depth of 5 and minimum samples split of 2, the model achieved a test accuracy of 97.5% when using all features and selected features.
- Similarly, with a maximum tree depth of 10 and minimum samples split of 3, the test accuracy improved to 98.6%. These results suggest that the model performed well in capturing the underlying patterns in the data and making accurate predictions.
- The test accuracy varied when no specifications were provided in instantiating the model, resulting in different accuracies such as 98.7% or 99.1%. This variability could be attributed to the randomness inherent in the algorithm, as well as the specific configuration of the model during training.

The heat-map confusion matrix in *Figure 6* shows the level of accuracy of the model with (a) untuned parameters and (b) tuned parameters.

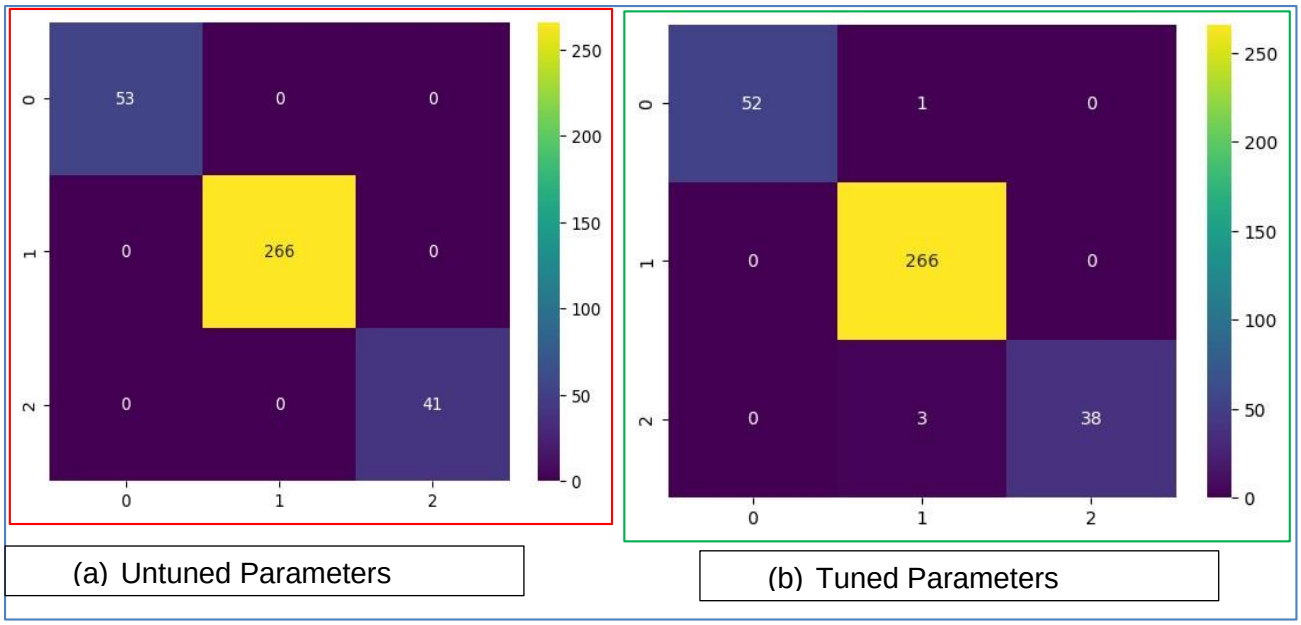


Figure 6: Confusion matrix for the Random Tree Classifier Model

From the confusion matrix above, the observations made are:

- Both models shows a high level of accuracy, with only a few misclassifications for the tuned option.
- Class 1 prediction (which can be compared to a Performance Rating of 2) gives 53 and 52 instances for the untuned and tuned parameters, respectively; 266 instances correctly predicted as class 2 (Performance rating of 3), and 41 vs 38 instances correctly predicted as class 3 (Performance rating of 4).
- There is one instance misclassified as class 2 and three instances misclassified as class 3 in the tuned model.

Despite the variability, the model consistently achieved high accuracy (>98%), indicating its robustness and generalization capability.

4.6 Model Deployment

To deploy the Random Forest Classifier model above, a Streamlit web application (INXapp.py) was created to demonstrate the analysis. The app development is based on:

- Streamlit (version 1.33.0) executed via its web-app from Anaconda Command prompt.
- The user is expected to load the raw data (.xls file) and thereafter the train dataset (.csv file) for training the model.
- The app also visualizes charts and tables for the univariate data (which can be expanded to cover more features).
- The sidebar and instructions are issued to the user and can be accessed vis the local host.

The app created allows users to upload a dataset, train a Random Forest classifier, and display evaluation metrics including the correlation and confusion matrices. A user can then further customize the app by adding additional features, such as, hyper-parameter tuning, or additional interactive visualizations. Model deployment will be done once the files are uploaded on github. See screenshots of the app in ***Appendix 2***.

5. Recommendations

The analysis gives an overall view into the employee performance, with the HR policies on salary increment and environmental satisfaction reflecting on better performance rating. Improvement policies would include:

i. Since the longer an employee has served in a department, the less the performance rating, inter-job-role change opportunities to those employees might improve their performance.

ii. The CEO should consider re-evaluating the key performance indicators (KPIs) and metrics for longer serving employees and those who have taken longer to get promotions. This will help track progress over time and adjust their day-to-day operations, assuming a monotonous work environment.

iii. The investment in employee learning and development seems to have almost no effect on the performance. A change in courses offered (with an input from employees) and creation of initiatives to enhance skills, competencies, and job satisfaction can contribute to overall organizational performance.

iv. Lastly, job rotation with different managers or co-workers can be used to gauge the nature of work periodically, since the longer employees are with one manager, the poorer the performance rating.

6. References

[1] A. Kimar, "A Quick Guide to Bivariate Analysis in Python," 27 November 2023 . [Online]. Available: https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/.

[2] M. Stojiljković, "Real Python," [Online]. Available: https://realpython.com/numpy-scipy-pandas-correlation-python/#heatmaps-of-correlation-matrices. [Accessed 12 April 2023].

[3] K. K. Al-jabery, "Computational Learning Approaches to Data Analytics," *Elsevier,* pp. 7-27, 2020.

[4] K. N. Jarapala, "Categorical Data Encoding Techniques," Medium, 14 March 2023. [Online]. Available: https://medium.com/aiskunks/categorical-data-encoding-techniques-d6296697a40f.

[5] K. Menon, Simply Learn, 15 Feb 2024. [Online]. Available:https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning.

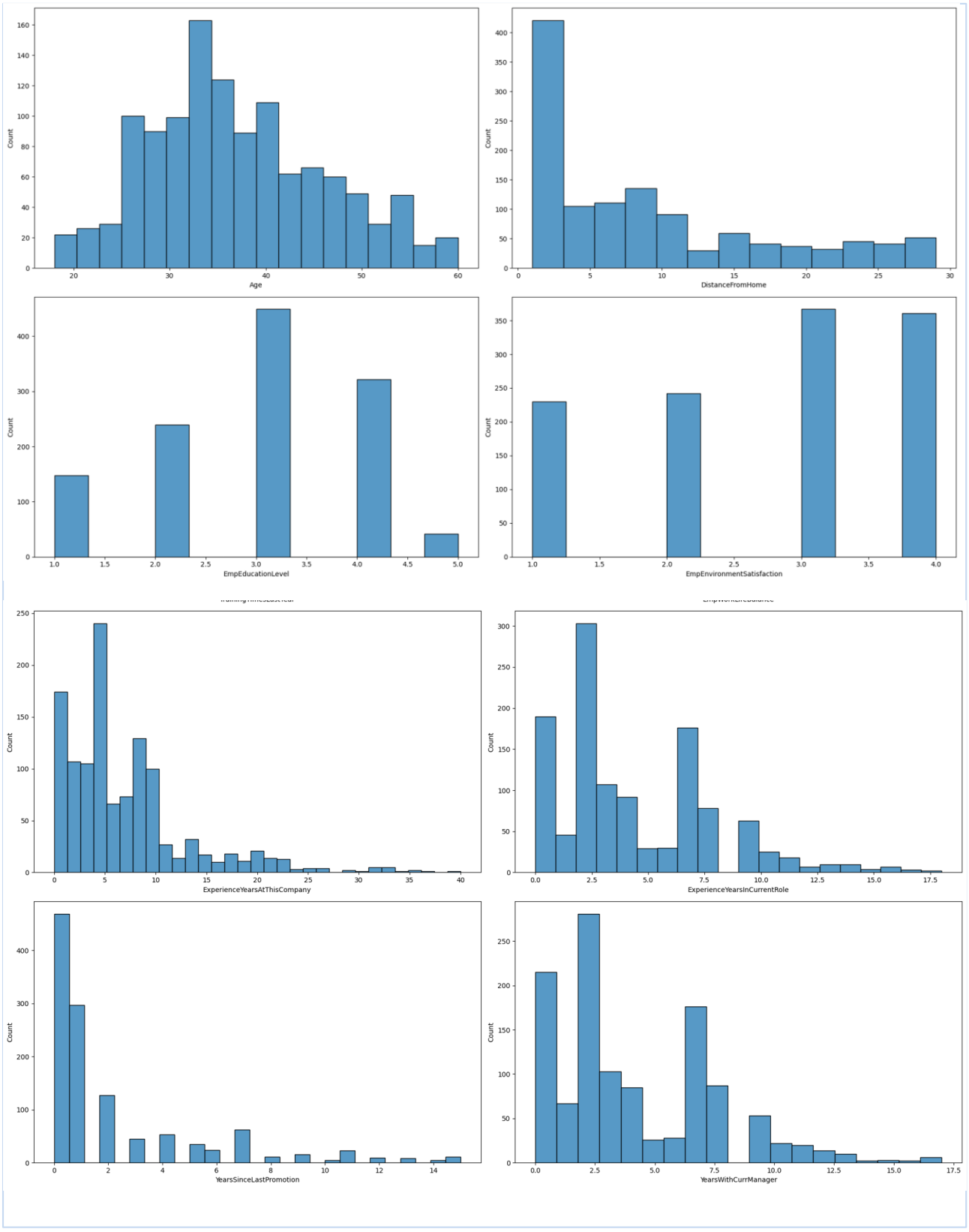
[6] A. AK, "Clever Cuts: Uncovering the Power of SelectKBest for Feature Selection in Machine Learning," Medium, 16 October 2023. [Online]. Available: https://medium.com/@abelkuriakose/clever-cuts-uncovering-the-power-of-selectkbest-for-feature-selection-in-machine-learning-c8d20d75c82f/

[7] J. Gitau, "July-ML-Classes," 2024. [Online]. Available: https://github.com/josephgitau/July-Ml-classes.

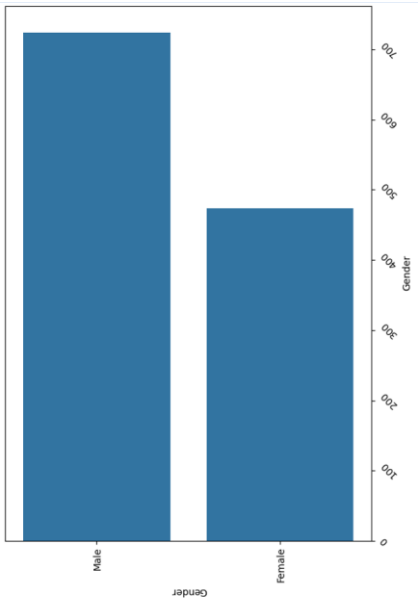
[8] A. Sarangam, "Classification vs Regression: An Easy Guide in 6 Points," 4 March 2021. [Online]. Available: https://u-next.com/blogs/artificial-intelligence/classification-vs-regression/

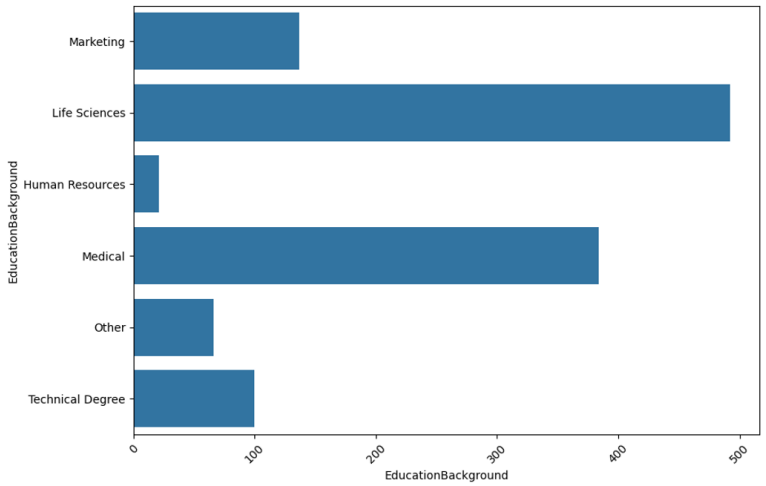
Appendix 1

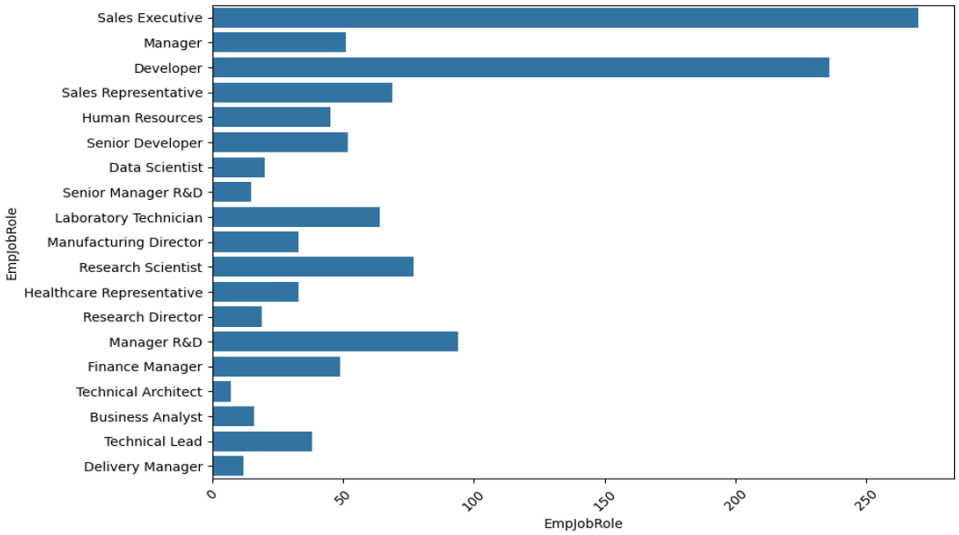
1a: Univariate Analysis - Sample Numerical Features

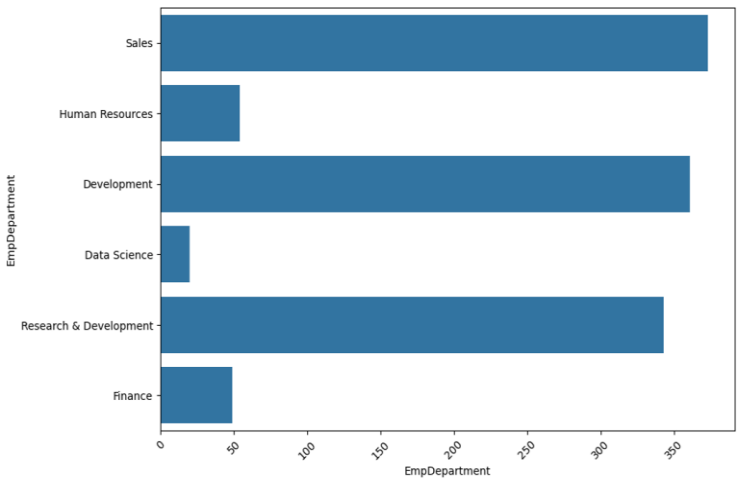


1b: Univariate Analysis - Sample Categorical Features









Appendix 2

Web App Development on Streamlit (sample screenshots)