Employee Performance Analysis

Project Code - 10281

INX Future Inc.

Candidate Name	:	Kapse Chandra Prakash
E-mail id	:	kcprakash778@gmail.com
Assessment ID	:	E10901-PR2-V18
Module	:	Certified Data Scientist - Project
Submission deadline date	:	18-Sep-2019
Registered Trainer	:	Mr. Ashok Kumar A
REP Name	:	DataMites™ Solutions Pvt Ltd
Project Assessment	:	IABAC™

Summary

The project aims to determine the factors which affect and weighs into the performance of an employee of INX Future Inc. . The dataset is provided of 1200 records of employees and 18 features, all factoring into their performance rating. The object of the project was to determine :

1. Department Wise Performance
2. Most important features affecting the employee performance rating.
3. Trained model to predict the performance rating for current employees and for future hiring.
4. Performance improvement recommendations.

In order to achieve these results we have to complete some preliminary operations which is imperative in order to meet statistical requirements.

The features of the data is classified into Numerical and Categorical features. They are as follows:

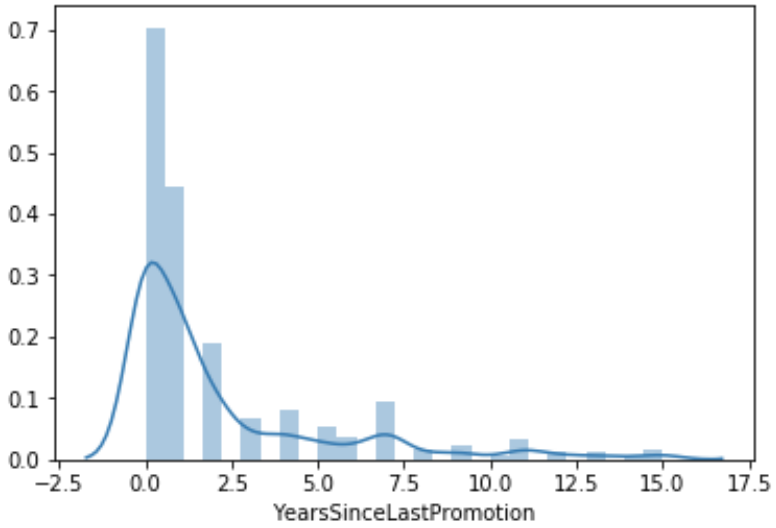
Numerical Features	Categorical Features
• Age	• Gender
• DistanceFromHome	• EducationBackground
• EmpHourlyRate	• MaritalStatus
• NumCompaniesWorked	• EmpDepartment
• EmpLastSalaryHikePercent	• EmpJobRole
• TotalWorkExperienceInYears	• BusinessTravelFrequency
• TrainingTimesLastYear	• EmpEducationLevel
• ExperienceYearsAtThisCompany	• EmpEnvironmentSatisfaction
• ExperienceYearsInCurrentRole	• EmpJobInvolvement
• YearsSinceLastPromotion	• EmpJobLevel
• YearsWithCurrManager	• EmpJobSatisfaction
	• OverTime
	• EmpRelationshipSatisfaction
	• EmpWorkLifeBalance
	• Attrition
	• PerformanceRating

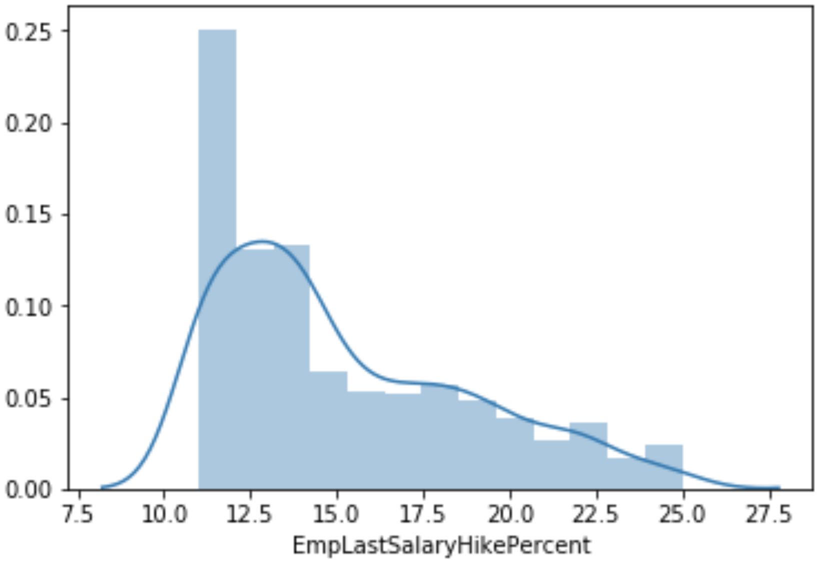
Data Cleaning

The data is checked for missing values of the features. The dataset provided to us has no missing values.

Visual Analysis

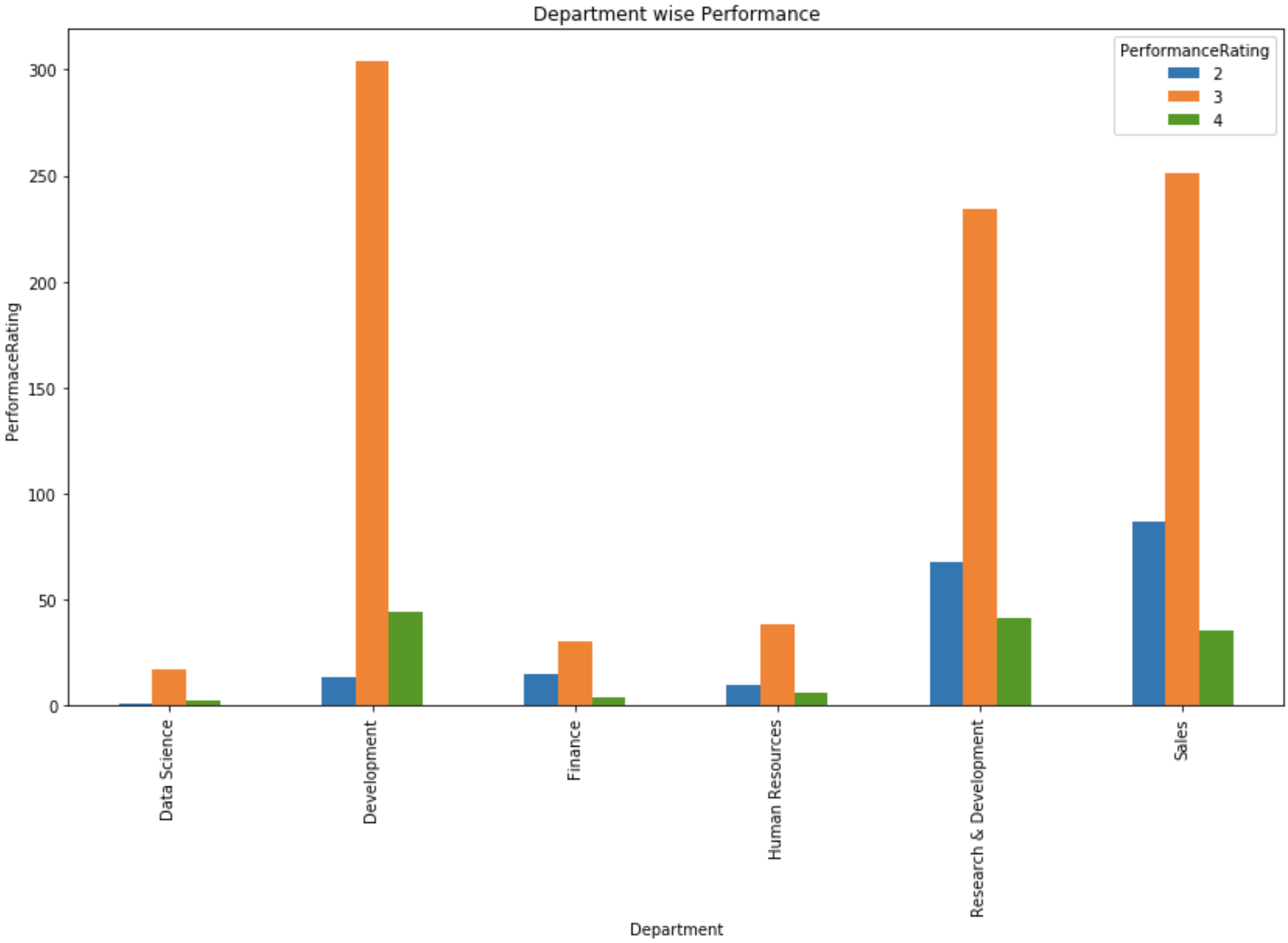
The visual analysis of various features helps in UNDERSTANDING the data or extracting insights from the data. The first and foremost visual analysis to be done on the data is to check whether the numerical features are normalised.





From the data, we can identify that the YearsSinceLastPromotion and EmpLastSalaryHikePercent features are negatively skewed and most likely to have an outlier

From the raw data the first objective can be found and plotted on bar graph as shown below.



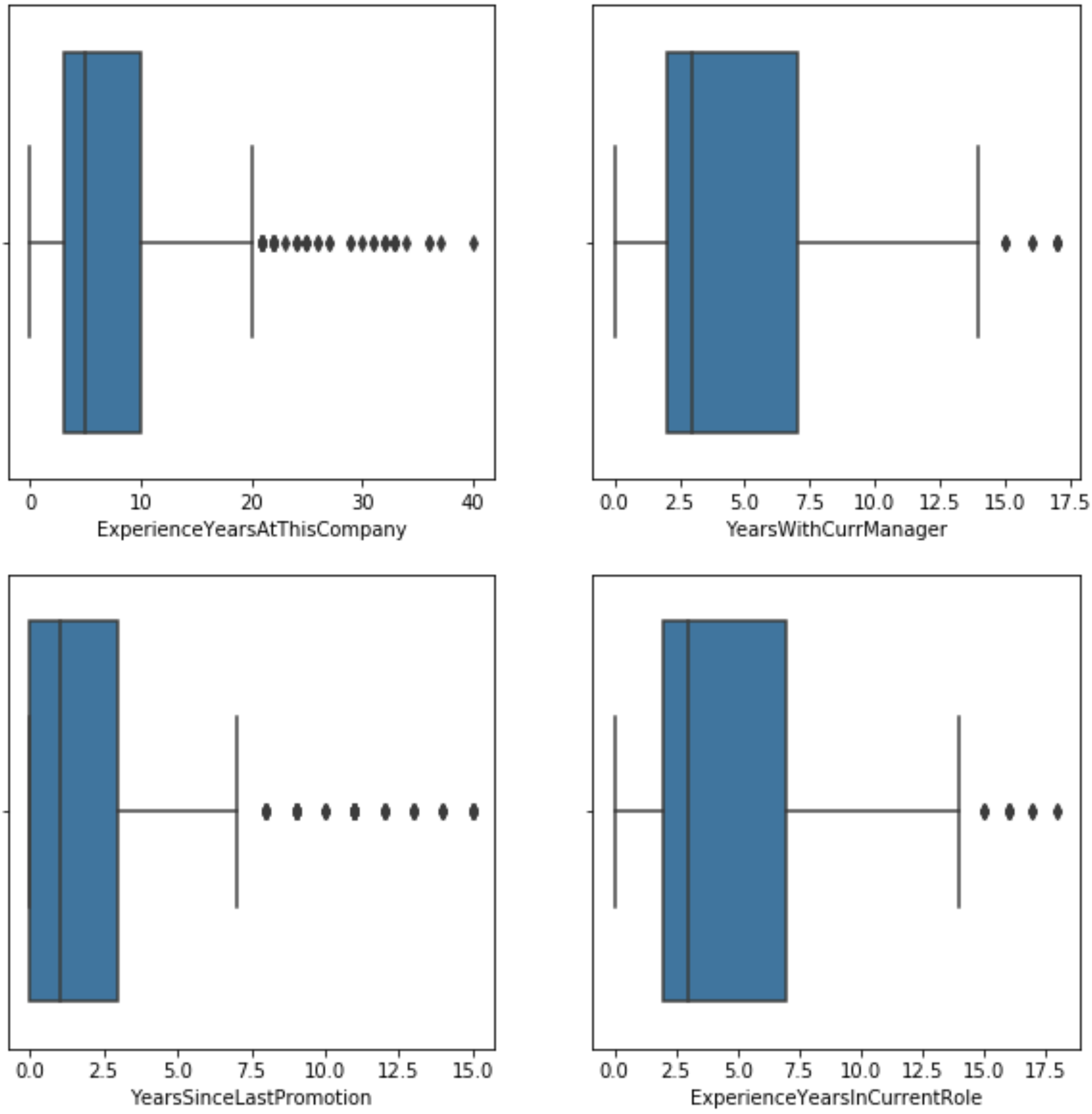
From the graph it is easily concluded that the mean rating of employee of every department is 3 and the highest number of employees having 3 rating are from R&D and Development Department.

Let's better understand the data more deeply by outlier detection , but as the dataset has 18 features, it becomes tedious to find outlier for every feature and of which some are categorical also, so to avoid this we first find out the most important features of the dataset by using SelectKBest score package from scikit-learn library.

This package tests the statistical relation of the feature with the target feature and can give the weight of the number of features as desired by the user. I used it to find top 5 most important features which affects the target. They are shown below with their SelectKBest Scores:

EmpEnvironmentSatisfaction	297.136917
EmpLastSalaryHikePercent	238.004284
NumCompaniesWorked	133.602068
OverTime	120.860036
MaritalStatus	110.691319

As these are the important features, outlier analysis can be done on them to further increase the efficiency of the model



Here we used Boxlpots from the Seaborn library to find the outliers and then treating these outliers with their median values as the whole dataset's numerical features has discreet values and continuous values for several features won't make any sense.

Machine Learning Model

The models used for prediction for this dataset are:

1. RandomForestClassifier
2. XGBClassifier

RandomForestClassifier

This is an ensemble algorithm which creates a set of decision trees of random features and predicts the output. The advantages of this algorithm is it eliminates the problem of bias and has good accuracy and works well with small as well as large datasets.

A little hitch - As per the observation of target records the number of performance rating of 3 category is higher than others which may push the output to 3 for majority of input. This is known as imbalance dataset. To avoid this we use SMOTE(Synthetic Minority Over-sampling Technique).

SMOTE creates synthetic samples of minority class from the similar records in the dataset.

Instead of splitting the data set into training and testing and passing it into algorithm , I used cross val score which divides the training and testing data into 5 folds and calculates efficiency for all the folds.

After this simply using GridsearchCV and passing on the RandomForestClassifier's parameters to this to find the best parameters for the algorithm.

The efficiency of the model was found to be - 95.6%

XGBClassifier

This is a boosting algorithm, as the name itself suggest that it boosts the learning from weak to strong learning sequentially. This is a very fast algorithm and works by fitting the model from the residuals of previous iteration.

Following the aforementioned methods, the efficiency of the model was found to be - 92.6%.

Results and Insights

- At around 100 employees in Sales have rating of 2 and w.r.t to Data Science department they have a good overall employee rating.

- The age group of 30-40 experiences low environment satisfaction and also declines in the performance as well.

- There has been no salary hike of employees having experience of 27-35 years in Human Resources Department and number of people with rating 4 w.r.t to this department is higher than other.

- The lowest number of employee with good overall department performance is development department and with significant salary increment.

- The finance department experiencing a poor performance and environment satisfaction for the age group of 25-27 years.

According to SelectKBest score the top 3 features affecting the employee performance are:

- Environment Satisfaction
- Salary Hike
- Number of Companies worked

and if these could be improved the employee performance also experience the same. Further the work experience when compared to number hours worked should be little lenient.
The experience of employee should to be taken into account when promoting or hiring.