PROJECT : EMPLOYEE PERFORMANCE ANALYSIS

NAME : HARSHA M N

ASSESMENT ID : E10901-PR2-V18

INDEX

FIGURES

1. PROJECT SUMMARY

In this project the main aspect is to predict the employee performance based on the data given. In the given data there are 1200 rows and 28 columns of data and Employee performance is graded as 2, 3 and 4. These are the grades given to the performance of employees.

The algorithm used for the prediction is an ensemble technique i,e. Random Forest. OOB method is used to increase the efficiency and also K-fold technique is used to find the average efficiency of the model.

1.1 Algorithm and Training methods used

The best algorithm that predicts very efficiently by taking minimum number of features for prediction is Random Forest.

Multivariate Logistic regression also can be used to predict the employee performance but it requires minimum of 20 features to predict, even then efficiency that can be obtained is 94% which acceptable but when considered the number of features to predict this model does not seem to be efficient.

1.2 Important Feature selected for analysis

In techniques such as PCA analysis the first principle component captured only 66.6% of the efficiency and the rest 33.3% is almost equally distributed among other variables as given the figure below

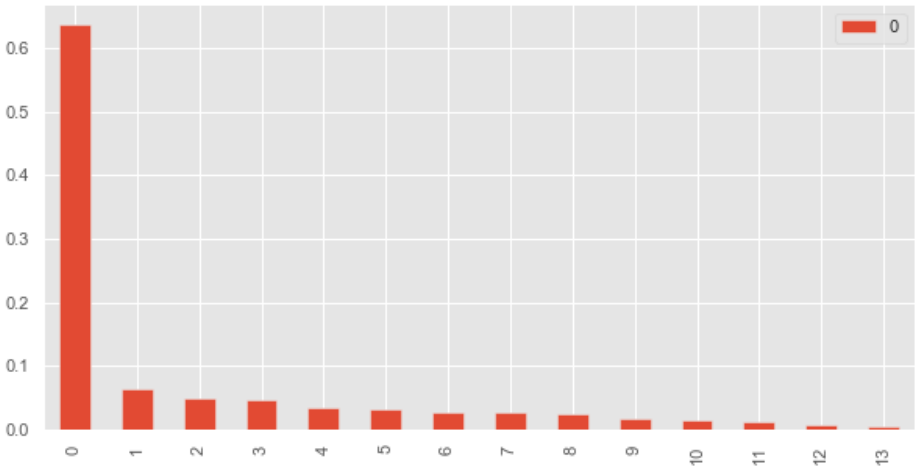

fig 1.1

In the above figure we can clearly observe that even after considering PC2 only 70% of data is captured and rest of the data is almost equally captured among the rest of the components. This makes the PCA analysis to reduce the features less efficient. So, PCA is not used for the analysis.

The features that is used for the analysis after feature reduction using correlation are

- Age
- DistanceFromHome
- EmpHourlyRate
- NumCompaniesWorked
- EmpLastSalaryHikePercent
- TotalWorkExperienceInYears
- YearsSinceLastPromotion
- YearsWithCurrManager
- EducationBackground
- EmpJobRole
- EmpDepartment
- EmpEnvironmentSatisfaction
- OverTime
- EmpWorkLifeBalance

In Random Forest, Grid Search is used to find the minimum features that can be used to predict the model.

Parameters obtained from the Grid Search are,

- Bootstrap : True
- criterion : entropy
- max_depth : 4
- max_features : None

Instead of using ‘Gini’ as the splitting parameter for the tree we are using ‘Entropy’ and the tree is limited to a depth of only 4 and Bootstrap aggregation is used randomly choosing data for the tree. Using these parameters the features that are capturing the data for prediction are,

- DistanceFromHome
- EmpDepartment
- EmpLastSalaryHikePercent
- YearsSinceLastPromotion
- YearsWithCurrManager
- EmpJobRole
- EmpEnvironmentSatisfaction
- EmpWorkLifeBalance

In the below figure it is shown how all the data is captured by only these 8 features

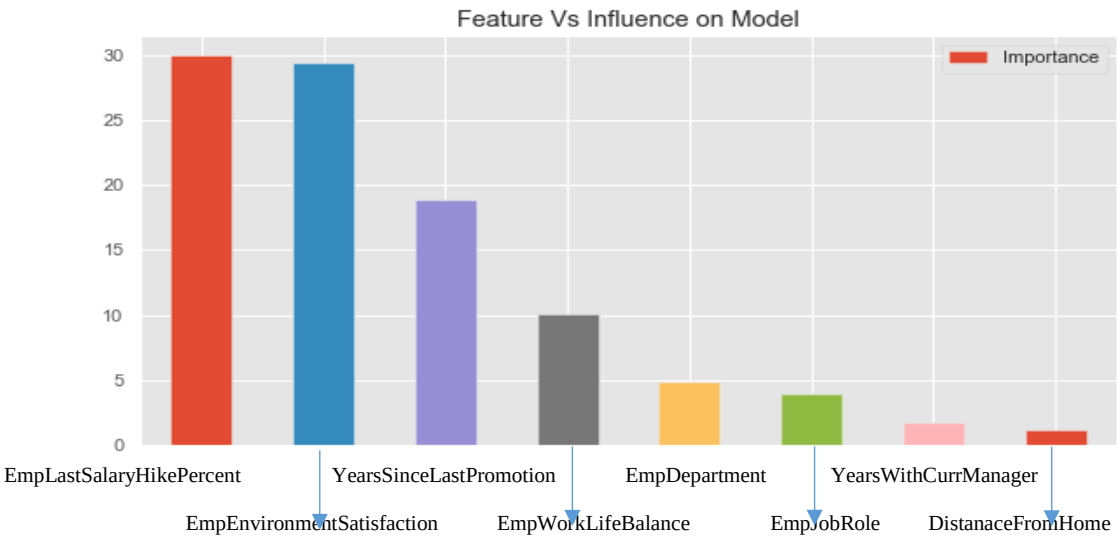

fig 1.2

By applying Grid Search for tuning the parameters the best accuracy score obtained is 95.55%. By considering these features the misclassifications that are happening is less.

The miss classification using cross tab is shown below,

Array ([[24, 2, 0],
[3, 132, 1],
[1, 1, 16]], dtype=int64)

1.3 Other Techniques used

Other techniques that is used to increase the efficiency of the model is OOB (Out Of Bag error). Usually while the data is split some of the data is not considered for training the data. These samples are called out of bag samples and the error calculated on these samples is called Out Of Bag error. Using this sample size of the tree can be decided.

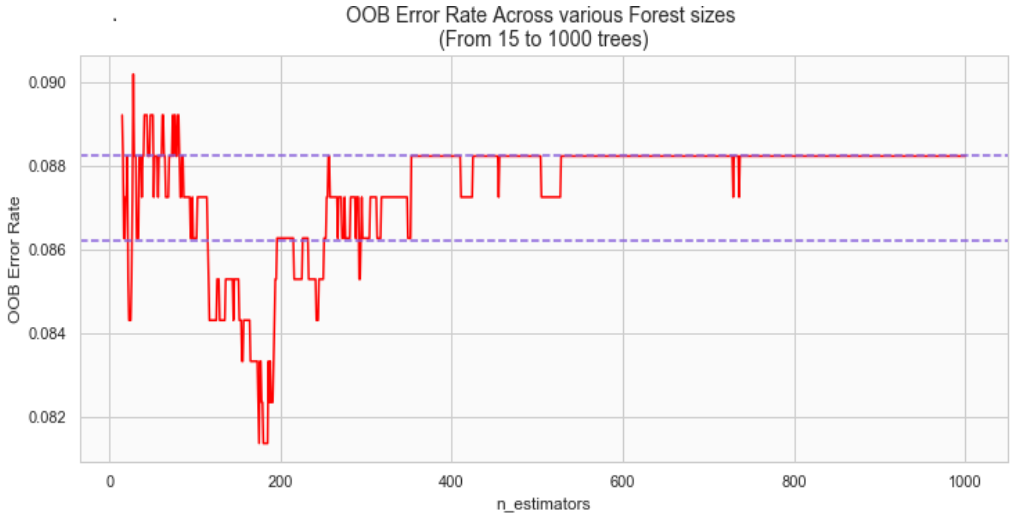

fig 1.3

In the above figure it is observed that around 300 tree samples it starts to oscillate so the n_estimator value is taken as 300 and **error rate** for 300 trees is **0.08627.**
After setting n_estimator parameter as 300 the efficiency increased to 96.66%

2. FEATURE SELECTION/ENGINEERING

There were no null values found in the given data, no special characters and no missing values and the given data is a mixture continuous and categorical value and predictor variable is a categorical value. The given data is clean.

2.1 Important Features selected for analysis

The given data is divided into categorical and continuous variable and based on the correlation values the 8 features were selected for prediction. They are,

- DistanceFromHome
- EmpDepartment
- EmpLastSalaryHikePercent
- YearsSinceLastPromotion
- YearsWithCurrManager
- EmpJobRole
- EmpEnvironmentSatisfaction
- EmpWorkLifeBalance

2.2 Feature Transformation applied

The continuous variables are

- Age
- DistanceFromHome
- EmpHourlyRate
- NumCompaniesWorked
- EmpLastSalaryHikePercent
- TotalWorkExperienceInYears
- YearsSinceLastPromotion
- YearsWithCurrManager

Normality of the continuous data checked and it is found to be non-normal as shown in the graph below

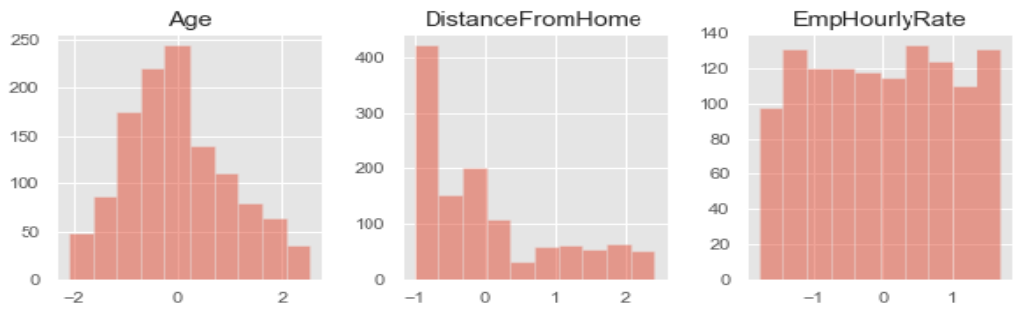

fig 2.1

So, Normalization of the data is done by applying scale.

Boxplot is done to find the outliers present which affects the prediction model

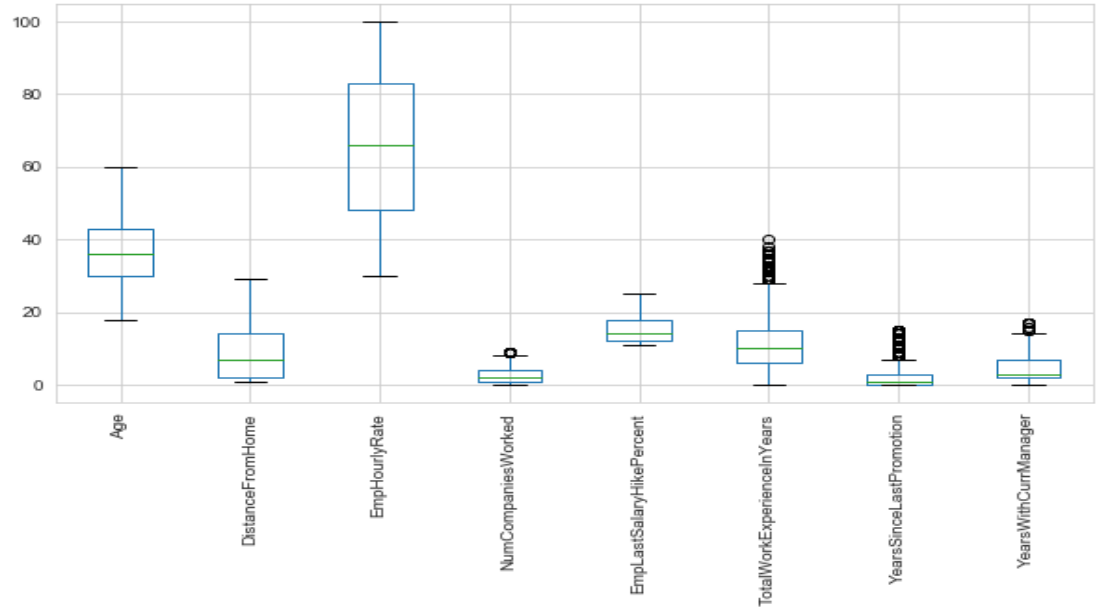

fig 2.2

In the above figure the data which is out of the Inter Quartile range should not be considered as outliers as they are only relatively far from the IQ range.

The categorical variables are

- EducationBackground
- EmpJobRole
- EmpDepartment
- EmpEnvironmentSatisfaction
- OverTime
- EmpWorkLifeBalance
- Gender
- MaritalStatus
- BusinessTravelFrequency
- EmpEducationLevel
- EmpJobInvolvement
- EmpJobLevel
- EmpJobSatisfaction
- EmpRelationshipSatisfaction
- TrainingTimesLastYear
- Attrition

Label Encoding was applied to all the categorical data.

Outcome variable is 'EmpPerformaceRating'

2.3 Correlation performed to select variables

Pearson correlation method was applied on the continuous variable and was mapped on heat map as shown below.

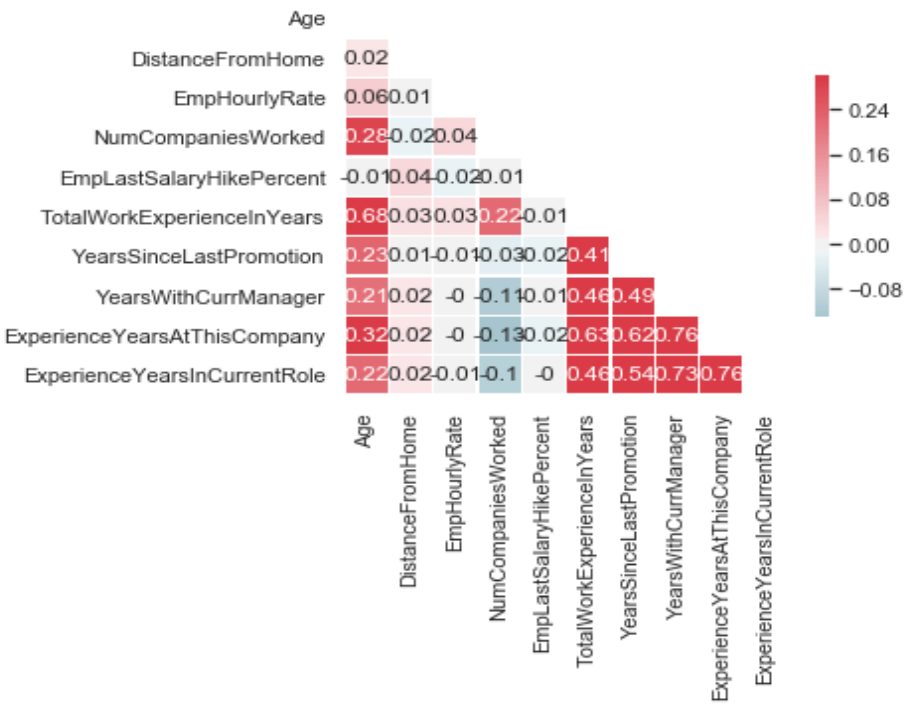

fig 2.3

In the above figure we can see the correlation values and the column having corr > 0.7 is removed.

- ExperienceYearsAtThisCompany
- ExperienceYearsInCurrentRole

are removed .

Chi Square test was used to remove the categorical columns on EmpPerformanceRating. The columns having p-value > 0.05 is removed and rest is considered for prediction.

- EducationBackground P-value : 0.4390094778578538
- EmpJobRole P-value : 1.3119815209234793 e-07
- EmpDepartment P-value : 3.832079093105312 e-11
- EmpEnvironmentSatisfaction P-value : 2.7264598025505243 e-67
- OverTime P-value : 0.0035485181432246523
- EmpWorkLifeBalance P-value : 7.960025516517787 e-05
- Gender P-value : 0.9217557495859275
- MaritalStatus P-value : 0.18941041088738678
- BusinessTravelFrequency P-value : 0.35490252129639344
- EmpEducationLevel P-value : 0.3114744353710753
- EmpJobInvolvement P-value : 0.9399863300500411
- EmpJobLevel P-value : 0.1768674128646029
- EmpJobSatisfaction P-value : 0.06534321929959395
- EmpRelationshipSatisfaction P-value : 0.8237615806491588
- TrainingTimesLastYear P-value : 0.8230699559751452
- Attrition P-value : 0.2770534777511834

The column EmpJobRole, EmpDepartment, EmpEnvironmentSatisfaction, EmpWorkLifeBalance, are considered and rest all removed.

3. RESULTS ANALYSIS AND INSIGHTS

3.1 Important Technique used in this project

The most important technique used in this project is the **K-fold** validation technique. In Random forest there are many random states which are created and it is picked by the system t hrough program but for each different random state we get different accuracy. So, in order to know the average efficiency of the model K-fold validation is done.

The K-fold validation accuracy is **0.913 (+/- 0.014)**

3.2 Answers to the problems mentioned in the project

1. Department wise performance

In the given below figure department wise performance is shown. It can be seen that Develop ment department has the highest rating and also highest average performing employee are pre sent in Development team only, Data Science field has no employee who has got rating as '2' Which implies that all the employee in this field are good performers.

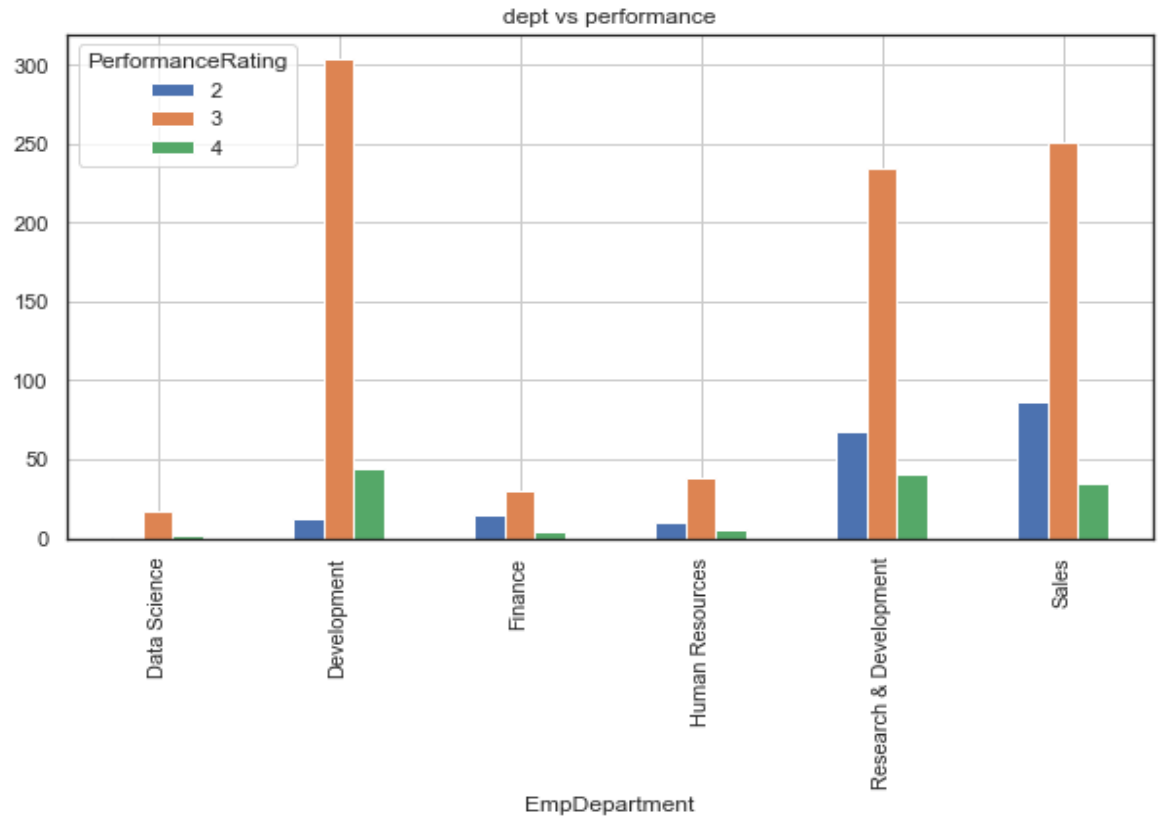

fig 3.1

2. Top 3 features affecting employee performance

From the below graph the top 3 features affecting the employee performance can be seen

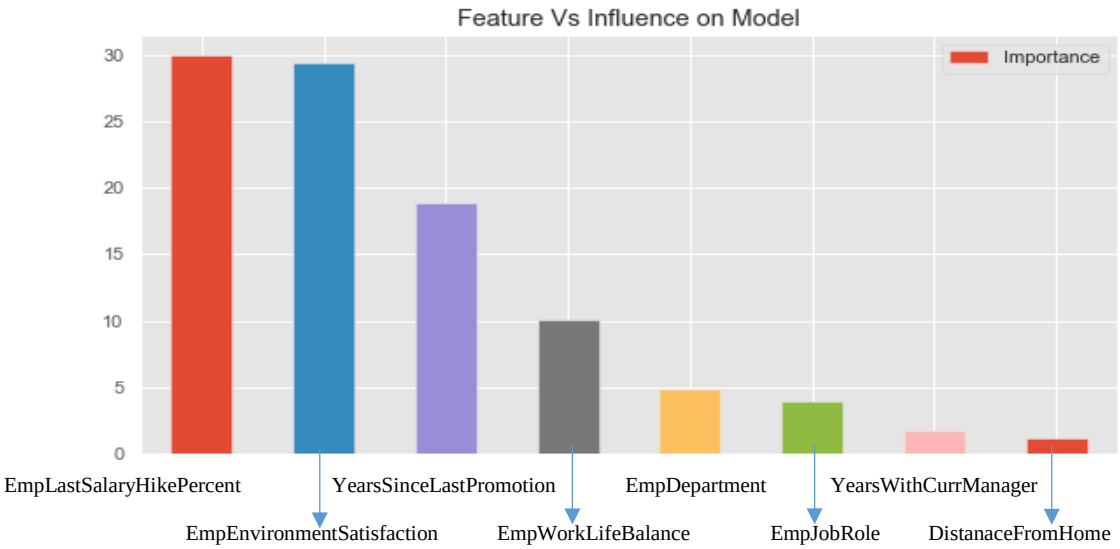

fig 3.2

a) EmpLastSalaryHikePerecent 29.612%
b) EmpEnvironmentSatisfaction 28.971%
c) YearsSinceLastPromotion 18.240%

3. A trained model which can be used to predict the employee performance based on fact ors as input has been attached in the zip file which can be used for hiring employees.

4. Recommendations based on insights

Attrition of people who have less experience is more it can been seen in below graph

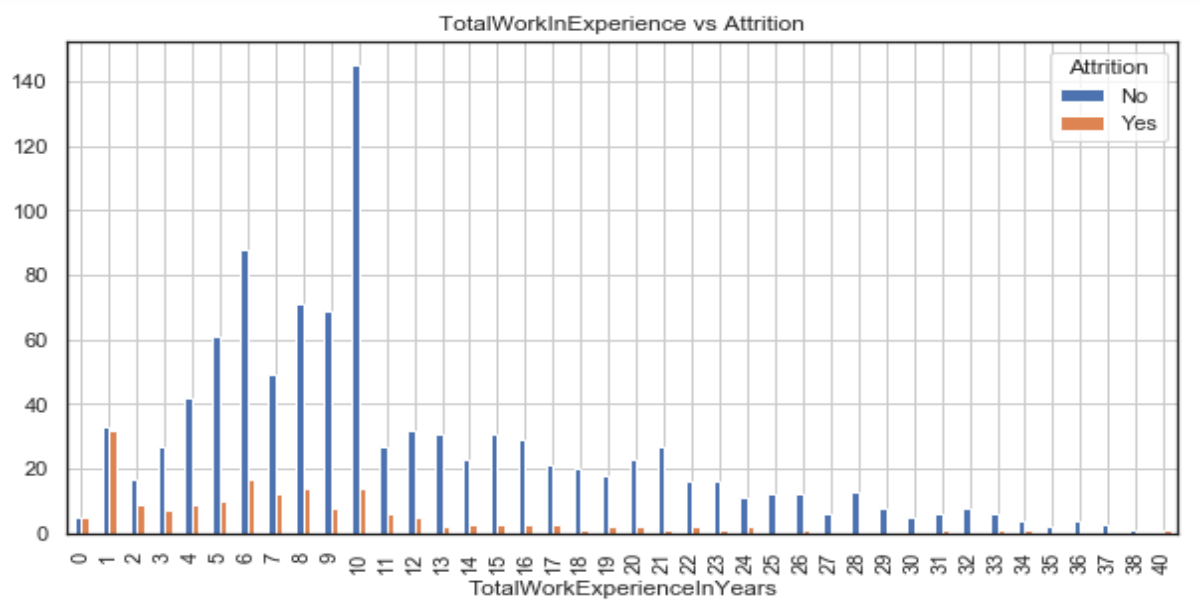

fig 3.3

Attrition of people is more in Development and Sales department it is shown in below graph

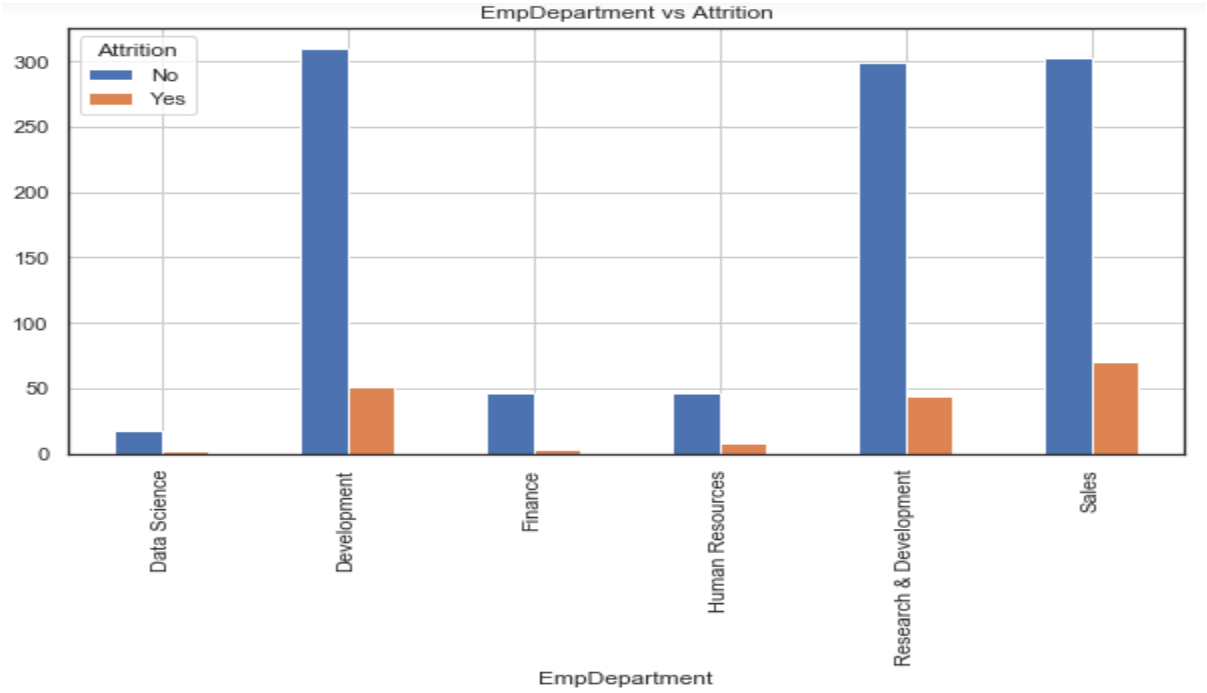

fig 3.4

Attrition of these people will not affect the moral of other employees because, job involvement of the people in the two mentioned department is less and also more inexperienced employees are present in these two department it can be observed in the below two graphs. There are more people who are graded 1 and 2 for their job involvement.

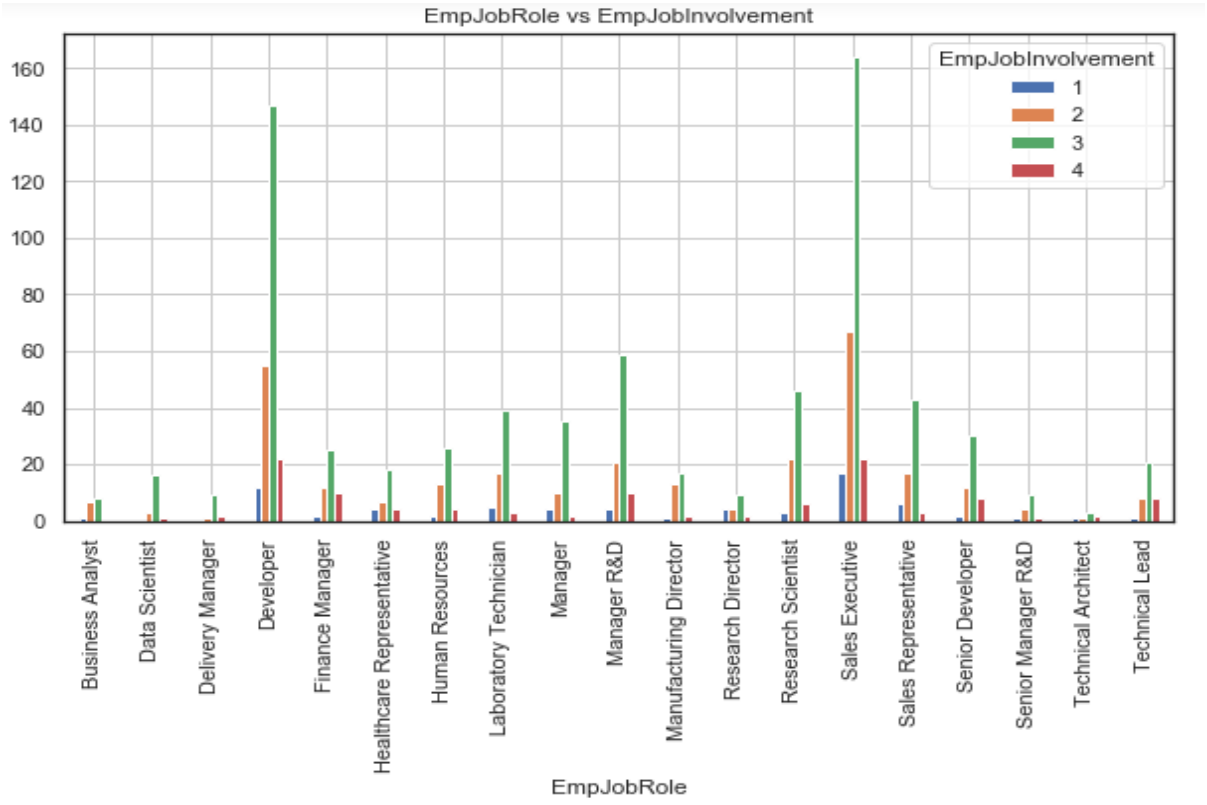

fig 3.5

From the below figure it is observed that more number of sales employee have 0-1 years of experience and more number of employees in Development have 5-7 years of experience which is less.

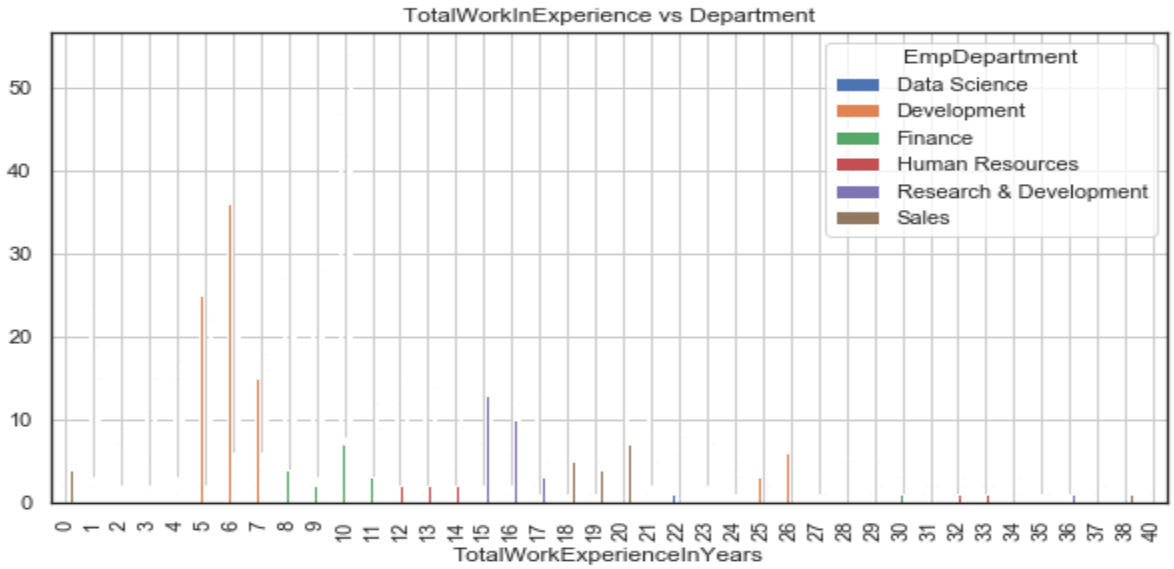

fig 3.6

So if more attention should be given towards Development and Sales team and regarding the employee performance if the employees are place in a more pleasant environment where they can work with peace of mind and if they are provided with proper salary hikes and then the promotion based on their experience the performance of the employees can be increased.

3.3 More insights from analysis

Attrition of employees is more for the employees who have less experience with the current manager they are working under. It can be seen in the below figure. As the experience with the current manager is increased there is no attrition of employees

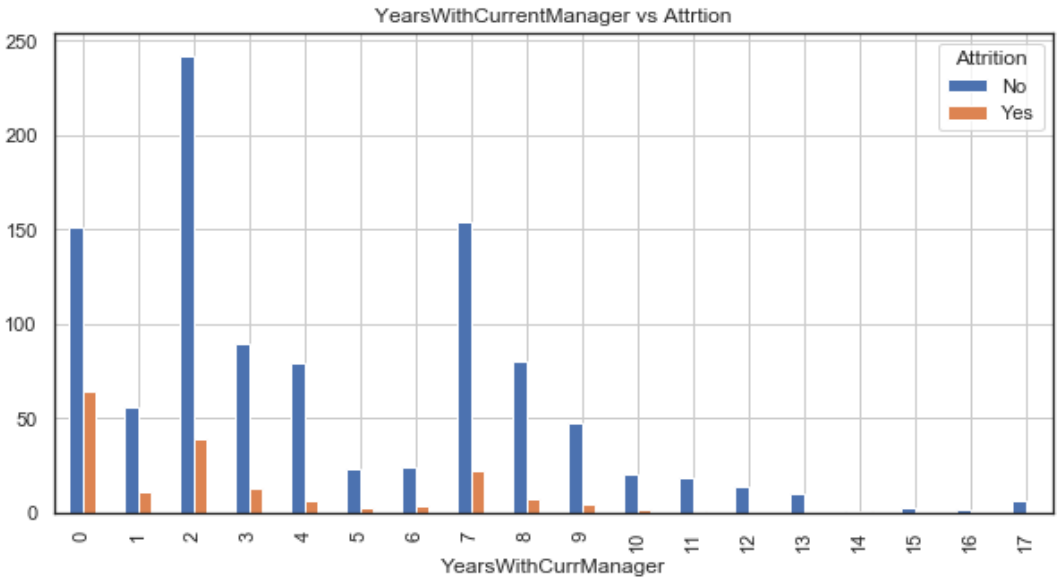

fig 3.7

More of the people who are working under the current manager and have less experience are from Development and Sales department. It is shown in below graph.

So it would be better if the manager for these two departments are changed.

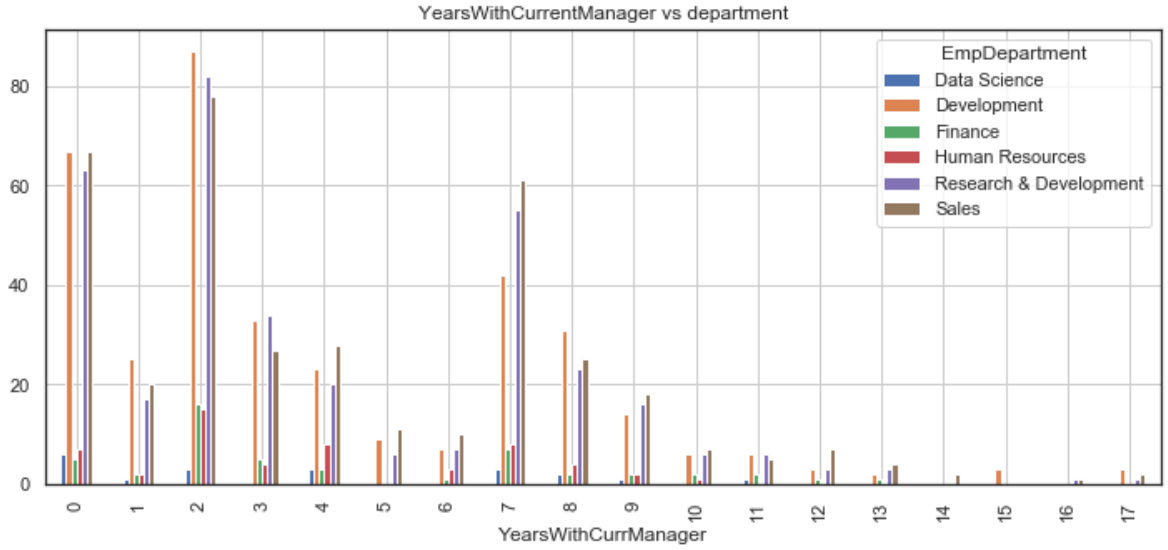

fig 3.8

REFERENCE

[1] www.mash.dept.shef.ac.uk/Resources/MASH-WhatStatisticalTestHandout.pdf · PDF file

For different type of correlation test.

[2] https://www.vaishalilambe.com/blog/data-science-algorithms-random-forest

For understanding random forest.