Employee Performance Analysis

INX Future Inc.


```
Candidate Name             : Ishan Pradip Borker
Candidate Email            : ishan123ppp@gmail.com
Assessment ID              : E10901-PR2-V18
REP Name                   : DataMites™ Solutions Pvt Ltd
Venue Name                 : Open Project
Exam Country               : India
Module                     : Certified Data Scientist - Project
Language                   : English
Exam Format                : Open Project- IABAC™ Project Submission
Submission Deadline Date   : 04-Apr-2020
Submission Deadline Time   : 23:59 Hrs [IST]
Registered Trainer         : Ashok Kumar A
Project Assessment         : IABAC™
```


Project Summary

1. 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. Mr. Brain, being a data scientist himself, expects the findings of this project will help him to take right course of actions. He also expects clear indicators of non performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.

The following insights are expected from 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 Jupyter notebook by using Python language.

2. Analysis

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

The categorical fields are **1. Gender, 2. EducationBackground, 3. MaritalStatus ,4. EmpDepartment, 5. EmpJobRole, 6. Business TravelFrequency ,7. Overtime , 8. Attrition**. These values are nominal, ordinal, ratio or interval.

The numerical fields are **1. Age, 2. DistanceFromHome, 3. EmpEducationLevel, 4.EmpEnvironmentSatisfaction, 5.EmpHourlyRate, 6. EmpJobInvolvement, 7.EmpJobLevel, 8.EmpJobSatisfaction, 9. NumCompaniesWorked, 10. EmpLastSalaryHikePercent, 11.EmpRelationshipSatisfaction, 12.TotalWorkExperienceInYears, 13.TrainingTimesLastYear, 14.EmpWorkLifeBalance, 15. ExperienceYearsAtThisCompany, 16. ExperienceYearsInCurrentRole, 17. YearsSinceLastPromotion, 18.YearsWithCurrManager, 19. 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 removing them.

Step 2: Perform Visualization(Graphical Representation) inorder 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.

Step 3: Checking for the outliers and removing it.

Outlier is a data point that differs significantly from other observations.

An outlier is caused due to variability in the measurement or experimental error.

It can cause serious problems in statistical analysis.

We use boxplot to find out if any outliers are present in any of the fields of the data.

Outliers were present in the following fields:

1. TotalWorkExperienceInYears
2. ExperienceYearsAtThisCompany
3. YearsSinceLastPromotion

After removing , we drop these fields and generated the corresponding new fields were generated as

1. clean_TotalWorkExperienceInYears
2. clean_ExperienceYearsAtThisCompany
3. clean_YearsSinceLastPromotion

Step 4: Define X and y variables.

Here X represents input variables and y represents output variable.

Step 5: Using Train-Test split ,scaling, 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 6: Machine Learning Algorithm to predict the employee performance.

Algorithms used are

- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- Artificial Neural Networks
- K- Nearest Neighbors
- Logistic Regression
- Support Vector machine
- Decision Tree Classifier

Steps for designing the machine learning algorithms

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

3. **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.

4. **Synthetic Minority Over-sampling Technique** (SMOTE) in Random Forest Classifier

Steps:
A. Import the required package.
B. Use train-test split.
C. Define and train the model.
D. Predict the model.
E. Display the confusion matrix and crosstab.
F. Calculate the accuracy score, precision score, recall score and F1 score.
G. Display the classification report. 8.Displaying X_train and y_train data after performing Synthetic Minority Over-sampling Technique(SMOTE).

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**.

3. 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

Also the analysis for Male/Female Performance Rating was done.

We can conclude that the highest average performance rating for Males is in '**Data Science**' Department and lowest is in '**Finance**' Department.

The highest average performance rating for Females is in '**Development**' Department and lowest is in '**Finance**' Department.

Overall Average Performance Rating of Females is more compared to Males.

Female --> 2.949474 Male --> 2.947586

Also analysis of each department is carried out with respect to Performance Rating index 2,3 4.

It was found that PerformanceRating 2 is highest in **Sales, Research and Development,Human Resources, Finance** departments.PerformanceRating 3 is highest in **Development and Data Science** department. PerformanceRating 4 is 2nd highest in **Development, Research and Development, Human Resources and Data Science** departments.

2. The Top 3 Important Factors affecting the employee performance

After the visualization and correlation matrix, it clearly indicates that the top 3 factors which affect the employee performance are

1) Employee Environment Satisfaction --> 39.5561%

2) Employee Last Salary Hike Percent --> 33.3722%

3) Years Since Last Promotion --> 16.7629%

3. A Trained Model which can predict the employee performance

The trained model which are designed using machine learning algorithm to predict the employee performance are listed below along with their accuracy score and precision score as:

The Machine Learning Algorithm which has the highest accuracy, precision and recall is '**Gradient Boosting Classifier**'. '**K-Nearest Neighbors**' algorithm has the lowest accuracy, precision and recall.

Sr.No.	Machine Learning Algorithm	Accuracy Score	Precision Score	Recall Score
1.	Random Forest Classifier	94.16%	94.03%	94.16%
2.	Gradient Boosting Forest Classifier	96.25%	96.22%	96.25%
3.	eXtreme Gradient Boosting(XGBoosting) Classifier	95.41%	95.39%	95.41%
4.	Artificial Neural Networks	80.41%	78.91%	80.41%
5.	K- Nearest Neighbors	75.41%	66.88%	75.41%
6.	Logistic Regression	83.33%	83.33%	83.33%
7.	Support Vector Machine	80.83%	80.83%	80.83%
8.	Decision Tree Classifier	92.08%	92.53%	92.08%

4. Insights

Here Visualization of each factor was carried out to check for the employee performance. The results were

1) BusinessTravelFrequency : 'Non-Travel' employees performed better.

2) DistanceFromHome : Employees staying 16km away from home performed better.

3) EmpEducationLevel : Employees having education level 3 is the highest.

4) EmpEnvironmentSatisfaction : Environment Satisfaction Level 3 performed better.

5) EmpJobInvolvement : Employees Job Involvement level 2 performed better.

6) EmpJobLevel : Employee Job Level 1 performed better.

7) EmpJobSatisfaction : Employee Job Satisfaction level 4 performed better.

8) OverTime : Employees working overtime performed better.

9) EmpLastSalaryHikePercent : Employees who received 21% hike in the salary has better performance.

10) NumCompaniesWorked : Employees who worked in 4 different companies performed better on the job.

11) EmpRelationshipSatisfaction : Employee having Relationship Satisfaction level 3 has performed better.

12) TrainingTimesLastYear : Employees who have been trained 3 times last year have performed better.

13) EmpWorkLifeBalance : Employees having Worklife Balance level 4 has highest performance rating.

14) ExperienceYearsAtThisCompany : Employees who have 23 and 34 years of experience in this company performed better.

15) ExperienceYearsInCurrentRole : Employees having 12 years of experience in the current role performed better.

16) YearsSinceLastPromotion : Employees who haven't been promoted since 13 years after their last promotion have performed better.

17) YearsWithCurrManager : Employees who spent 17 years with the current manager has better performance rating.

18) EmpHourlyRate : Employees working at the rate of 38 hours have performed better.

19) Attrition : Employees who haven't resigned from many jobs have performed better.

Note: Graphs for all these factors can be found from "**INX-Future-Inc_Employee_Performance_Analysis_by_IshanBorker.ipnb**" source code file.

5. 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 GridSearch Cross Validation in Random Forest Classifier,

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

- Using RandomizedSearch Cross Validation in Random Forest Classifier,

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

- Using Synthetic Minority Over-sampling Technique in Random Forest Classifier,

> Accuracy Score was 89.16% ,Precision Score was 90.27% and Recall Score was 89.16%.

Feature Engineering technique results in highest accuracy score, precision score and recall score. **GridSearch Cross Validation** and **RandomizedSearch Cross Validation** techniques result in lowest accuracy score , precision score and recall score.

Gradient Boosting Classifier results in highest accuracy, precision and recall. Also **Random Forest Classifier**, **eXtreme Gradient Boosting Classifier** and **Decision tree Classifier** results in more than 90% accuracy, precision and recall.

6. 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 inturn boost the confidence of the employees.
- The other factors like Experience Years In Current Role (14.76%) , Employee Work Life Balance (12.44%), Years With Current Manager(12.23%) also needs to be monitored carefully for better functioning of the organisation.
- Males need to work hard inorder to be in par with Females with respect to Performance.They need to improve in Human Resources and Finance Departments.
- Females need to work better in Sales and Finance departments.
- Finance Department needs to closely monitor their employees as both males and females have not performed better.

Answers to the following questions

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. Gradient Boosting Classifier
3. XGBoost Classifier
4. Artificial Neural Networks
5. K- Nearest Neighbors
6. Logistic Regression
7. Support Vector machine
8. Decision Tree Classifier

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

Ans) 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. Grid Search Cross Validation (CV) in Random Forest Classifier : It is the process of performing hyper parameter tuning in order to determine the optimal values for a model. The performance of the entire model is based on the hyper parameter values.
2. 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.

3. Synthetic Minority Over-sampling Technique(SMOTE) in Random Forest Classifier : It is an oversampling method and creates the synthetic samples of minority class.It aims to balance class distribution by randomly increasing minority class examples by replicating them.
4. 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.

Ans) 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, xgboost**.

4. Did you make any important feature transformations?

Ans) Yes, 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.

5. Correlation or interactions among the features selected and how it is considered?

Ans) Yes correlation was selected among the features using python code

corr = performance.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 the correlation matrix.

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

6. Did you find any interesting relationships in the data that don't fit in the sections above?

Ans) 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.

7. What is most important technique you used in this project?

Ans) 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%.