Employee Performance Analysis - INX Future Inc.

Data Science Project Report

Submitted by

Floriann Deepika Louis

Assessment ID	**:**	**E10901-PR2-V18**
REP Name	**:**	**DataMites™ Solutions Pvt Ltd**
Module	**:**	**Certified Data Scientist – Project**
Exam Format	**:**	**Open Project- IABAC™ Project Submission**
Registered Trainer	**:**	**Ashok Kumar A**
Submission Date	**:**	**23-May-2020**
Email	**:**	**florianndeepika@gmail.com**

Contents

1. Project Requirement

This Data science projects includes the analysis of the employee performance of INX Future Inc. INX Future Inc, (referred as INX), is one of the leading data analytics and automation solutions provider with over 15 years. Recent years, the employee performance indexes are not healthy and this is becoming a growing concerns among the top management.

i) Project Goal

Goal of the project is to find the performance rating of the employees with various features available in the provided dataset.

The following insights that are expected to be analysed 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

ii) Input Dataset

The input dataset was given by IABAC.

Link:
http://data.iabac.org/exam/p2/data/INX_Future_Inc_Employee_Performance_CDS_Project2_Data_V1.8 .xls

Dataset name: INX_Future_Inc_Employee_Performance_CDS_Project2_Data_V1.8-2

Description: This dataset is based on the INX Future Inc (referred as INX). It is one of the leading data analytics and automation solutions provider with over 15 years of global business presence.

Programming Language: Python in Jupiter notebook

2. Analysis Techniques applied

Below statistical techniques and algorithms were applied for the analysis

i) Algorithm and training method(s) used

- The dataset is labelled and has target so supervised machine learning algorithm was selected to train the model.
- Since the target variable is categorical – ordinal data and is a multi-class, classification algorithms was selected.
- Following algorithms were modelled and the data was trained

    1. Random Forest Classifier
    2. K Nearest Neighbours
    3. Decision Tree
    4. Support Vector Machine
    5. XG Boost
    6. Naïve Bayes
    7. Artificial Neural Network

- The supervised machine learning model Random Forest Classifier predicted with an accuracy of 92% and XG Boost with 93%.

ii) Feature selection techniques

Three techniques were used for selecting features for training the model

1. Correlation
2. Feature importance
3. Backward Elimination of variables

iii) Other techniques and tools used

There were mixed of both categorical and numerical data. So label encoding technique was used as a pre-processing step to convert the string categorical data to numerical data.

3. Feature Selection and Engineering

Analysis were based on the featured present in the dataset. Features which are past data are used to predict the employee performances.

- The features in the dataset had categorical and numerical data
- A provided data is structured and has targets. It is a 1200x28 data with 1200 rows and 28 features.
- The features are combination of quantitative and qualitative data.
- 16 features are numerical data and 11 features are categorical data
- 1 is alphanumerical data (Employee data) which can be neglected since it doesn't have any effect on the performance of the employee.
- Target variable here is the Performance Rating, which is a categorical – ordinal data

i) Most important features selected for analysis:

Following are the important features selected which contributed more for the target variable which is the employee performance

i) EmpLastSalaryHikePercent
ii) EmpEnvironmentSatisfaction
iii) YearsSinceLastPromotion
iv) EmpJobRole
v) EmpDepartment
vi) EmpHourlyRate
vii) ExperienceYearsInCurrentRole

ii) Feature transformations

- Label encoding technique was used to convert the string categorical variables to numerical features so that all the independent variables are of same form
- Standardization technique was used, but it had no effect on the increasing the accuracy of the model. Hence removed it
- SMOTE is applied in order to make the data balanced

iii) Correlation or interactions

3 techniques were used to confirm the right features were selected to model the data.

a. Correlation - The features which ranged from -1 to +1 were selected and heat map was used to visualize the correlation between the independent and dependent variables.
b. Feature importance – First 10 features which contributed more towards the dependent variable were selected by using pandas.
c. Backward elimination process – Backward eliminating of features were done to confirm once whether the selected features are good to train the model.

4. Machine Learning Model

Following steps were followed to model the machine learning algorithm

Step1: Importing necessary packages

- All the necessary packages were imported to build the model. Pandas and Numpy were used for data computation and data frame usage.
- Matplotlib and Seaborn packages were used for visualization purpose.
- Sklearn and xgboost were used for the Machine learning algorithms.

Step2: Data importing

Using Pandas the given dataset was imported as a data frame.

Step3: Basic checks

- Here as a first step the alphanumerical feature which is employee ID is removed since it has no effect on our target variable
- Shape of the data is identified as 1200 rows and 28 columns.
- There are no NAN values in the dataset
- The data types are properly given. No typecasting required.
- The provided data is already structured there is no much data cleaning required.
- The target variables are found to be imbalanced. Hence SMOTE technique is used to make the data balanced.

Step4: Exploratory Data Analysis

- The relationship between the independent variables and the dependent variables is analyzed using the correlation.
- Correlation coefficient are checked by the values got by corr..
- These are also done by heat map. The heat signatures show the level of correlation from 0 to 1
- From the visualization we can conclude that,
  - Total year of work experience is highly correlated with the job level
  - Age is also highly correlated with the total year of work experience

- Experience at the current company also plays a vital role in the number of years with the current manager, year since last promotion and experience in the current role

Step5: Define X and y variable

- Using backward elimination process the useful features were identified as X variables
  - ✓ EmpLastSalaryHikePercent
  - ✓ EmpEnvironmentSatisfaction
  - ✓ YearsSinceLastPromotion
  - ✓ EmpJobRole
  - ✓ EmpDepartment
  - ✓ EmpHourlyRate
  - ✓ ExperienceYearsInCurrentRole

- y variable is PerformanceRating

Step5: Train Test split

- Splitting the data into train data and test data. Here train size is taken as 70%(840) and test is 30%(360)

Step6: Define the model

- The classification algorithm are defined at this stage. The models used were,
  - ✓ Random Forest Classifier
  - ✓ K Nearest Neighbors
  - ✓ Decision Tree
  - ✓ Support Vector Machine
  - ✓ XG Boost
  - ✓ Naïve Bayes
  - ✓ Artificial Neural Network

Step7: Fit the model

- X train and y train data are fit into these models

Step8: Evaluation and predictions

- Various metrics like accuracy score, Classification report with precision, recall, f1 score and confusion matrix are evaluated.
- The machine learning model which gives high accuracy are,
  - ✓ Random Forest classifier: 92% accuracy
  - ✓ Gradient boosted Classifier: 93% accuracy

5. Results and Insights

After Performing exploratory data analysis and supervised machine learning algorithm the following insights were found.

i) Result 1: Department wise performances

Visualization method was used to analyze the performance of employees with respective to the employees department. Seaborn and matplotlib are used to visualize date to get more insights on this category. The Employee department feature was taken and this was separated department wise. There are 6 departments available in the employee department variable.

- **Development**
  - Number of employees working in this department is second highest (361 employee) when compared to others.
  - Employees with technical back ground of life science and medical work here
  - Even though the count of Female employees are less when compared to male employees they perform well.
  - More employees contributes to 3 rating in performance.
  - They also have good amount of people who have 4 rating.
  - Employee with rating 2 lie within average age of 24 to 39.
  - People with 2 rating in salary hike, environment satisfaction, promotion and job involvement.

- **Data Science**
  - They have least amount of people working. Total of 20 employees but contribute more in employees performance rating
  - Number of female employees are less than the male employees.
  - This department has a mix of employees with different education background
  - People with 3 rating are high here.
  - It is found that employees with age around 40 are the ones with 2 rating. Otherwise age is not a constraint here to perform good here.
  - The head count they have doesn’t affect on their performances.
  - There is more scope to improve employee performance when there is increased job involvement of individuals

- **Human Resources**
  - Total employees working here are 54.
  - All employees are of same educational background.
  - The performance of female employees outstand the male employee even though they are less in numbers.
  - Employees with 2 rating mostly lie between the age of 31 and 44.
  - Some good performers lack in Employee job involvement and employee relationship satisfaction

- **Research & Development**
  - They have large numbers of employees working here. Total of 343.
  - People with different education background work here.
  - Female employees are less but they have good performance rating.
  - This department shows increased 2 rated employees than 3.
  - There is lack in employee environment satisfaction

- **Sales**
  - They have highest number of employee of 373 working.
  - Even though most of the employees are from marketing educational background, it is found that they contribute less in performance rating matrix
  - 2 rated employees are more than 3.
  - Employees with more 2 ratings are from sales department when compared to others, they lack more in employee environment satisfaction and employee job involvement

- **Finance**
  - Number of employees working here are 49.
  - Male employees perform well here. It is also found that people with 2 ratings are increased in number here.
  - Employee salary hike and environment satisfaction has great scope to improve, because many people lack here

ii) Result 2: The Top 3 important features affecting the employee performance

Feature engineering was done to identify the important features which affect the employee performances. Using correlation coefficient and the feature importance in random forest algorithm in sklearn the 3 importance features in the dataset was identified.

- ✓ Employee Salary Hike Percentage
- ✓ Employment Environment Satisfaction
- ✓ Years Since the last Promotion

iii) Result 3: A Trained model which can predict the employee performance

The dataset was trained model using supervised machine learning algorithm with classification method with the accuracy score.

- ✓ **Random Forest classifier: 92% accuracy**
- ✓ **Gradient boosted Classifier: 93% accuracy**

iv) Result 4: Recommendations to improve the employee performance

Summarizing from the above analysis, following steps are recommended to maintain and enhance the performance of the employees.

- ✓ Even though the number of male employees exceeds female in count, it shows that female contributes more when compared to male. More female employees can be recruited to increase the performance of the company.
- ✓ The Employees contribute more in performance are Business Analyst than the Senior Manager R&D. Area where they lack contribution (EmpRelationshipSatisfaction and EmpWorkLifeBalance) to the company has to be enhanced to increase their performance.
- ✓ More employees who contributes 2 rating. EmpJobLevel is low for Research & Development and EmpLastSalaryHikePercent is low for Sales department. This has to be revised for both the department.
- ✓ EmpJobSatisfaction, EmpJobInvolvement and EmpRelationshipSatisfaction has to be improved to increase the performance of the employees.
- ✓ Ensure the employees get salary hike at regular intervals with job level promotion. This will increase motivation for an employee to contribute more towards the performance of the company
- ✓ Constant monitoring of employees working environment satisfaction. Bringing more infrastructure in their working environment. This may boost them up psychologically to spend more hours productively.
- ✓ Regular trainings can be conducted. Workshops or providing online courses which will enhance employees to update their domain knowledge and skills in the department they are working in.
- ✓ Conducting fun games at workplace at regular basis contributes to a good work life balance.