Candidate Name : TCHUYA PAYONG ADRIEN

E-Mail : adrien_payong@yahoo.com

Project Code : 10281

REP Name : DataMites™ Solutions Pvt Ltd

Assesment ID : E10901-PR2-V18

Module : Certified Data Scientist - Project

Exam Format : Open Project- IABAC™ Project Submission

Project Summary

INX Future Inc, is one of the leading data analytics and automation solutions provider with over 15 years of global business presence. In recent years, the employee performance indexes are not healthy and this has become a growing concern among the top management. The CEO Mr. Brain, decided to initiate a data science project, which analyzes the current employee data and find the core underlying causes of the performance issues. He also expects a clear indicators of non-performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.
The Goal and Insights of the project are defined as follows :

- Départements wise performances.
- Top 3 Important factors affecting employee performance.
- A trained model which can predict the employee performance based on factors as inputs.
- Recommendations to improve the employee performance based on insights from analysis.

To satisfy the project goal, we have gone through the following steps :

- We have imported the data in Jupyter Notebook, provided the shape and description of the data and checked out null value.

- Analysis of each department, distribution and correlation analysis.
- Label encoding method was used to convert the categorical text data into model-understandable numerical data.
- Split the data into test and train.
- The following algorithms were used : K-Nearest Neighbor, Support Vector Machine, Naive Bayes Bernoulli, K-Nearest Neighbors, Random Forest with GridSearchCV, XGBoost Classifier
- Random Forest with GridSearchCV has predicted the highest accuracy of 92% follow by XGboost with 91% accuracy.
- The important features were predicted using Random Forest Classifier and Gradient Boosting algorithms.

Requirement

The source of the data provide for this project is IABAC and it is based on INX Future Inc. which is one of the leading data analytics and automation solutions provider.

Analysis

1) Data preprocessing

- The pandas method read_excel () was used to load the data.
- We have provided the shape of the data. shape is a tuple that gives you an indication of the number of dimensions in the array : The Shape of our data is 1200x28
- The describe() method was invoked to provide some statistical data like mean, standard deviation of the dataset…etc….
- We have invoked info() method to provide general information like the size of the data… etc…
- The formula dataset.describe (include= [np.object]) was used to include only string columns in the dataset description.

- dataset.isnull ().sum() was used to check null value in the dataset : The given data is well structured and cleaned and there are no missing data present in the data.
- dataset.to_csv('INX_Future_Inc_Employee_Performance_Exploratory.csv',index=False) was used to save the data.

2) **Exploratory Data Analysis**

a) **Department wise performance**

The departments are : Sales, Human Resources, Development, Research and Development, Finance

b) Numerical and categorical features

b.1) **Numerical features**

Numerical data have the meaning as a measurement or they are a count. Examples of numerical data are the salary of an employee, the age of a person, and the number of houses a person owns. Numerical data can be further categorized into two types: *discrete and continuous. In our case, the numerical features are :*

- Age
- DistanceFromHome
- EmpHourlyRate
- NumCompaniesWorked
- EmpLastSalaryHikePercent
- TotalWorkExperienceInYears
- TrainingTimesLastYear
- ExperienceYearsAtThisCompany
- ExperienceYearsInCurrentRole
- YearsSinceLastPromotion
- YearsWithCurrManager

b.2) **Categorical features**

Typically, any data attribute which is categorical in nature represents discrete values which belong to a specific finite set of categories or classes. These are also often known as classes or labels in the context of attributes or variables which are to be predicted by a model (popularly known as response variables). These discrete values can be text or numeric in nature (or even unstructured data like images!). There are two major classes of categorical data, nominal and ordinal. In our case, the categorical features are :

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

c) **Distribution of numerical and categorical features**

c.1**) Distribution of numerical features**

- Majority of the employees are in the age group of 30–40
- 0 to 30 unit : It is a distribution of the distance from home to office.
- *Maximum number of employees work in up to 2 compagnies among the 8*
- For most of the employees, the hourly rate range is 65 to 95
- Many employees were promoted in the last 0–1.5 years and employees were regularly promoted

c.2) **Distribution of categorical features**

- There is more male employees than female employees
- We have six unique educational backgrounds among employees In the company
- we can count nineteen unique employee job roles
- For most of the employees, education level is 3
- job satisfaction is very high for the highest number of employees
- Most of the employees are not having attrition
- There are almost 30% of the employees doing overtime

c.3) **Analysis through visualisation**

Data visualisation refers to any way of presenting information so that it can be interpreted visually. Through it, we can make use of our ability to recognise relationships to draw out meaning from data. We have made use of statistical plotting graphs for visualisation and correlation method to visualize the correlation heatmap.

c.3.1) **Distribution plot**

In order to have an idea of how the features are distributed with one another, we have invoked several functions from seaborn library (like countplot, barplot, boxplot). Seaborn is a module in Python which is built on top of matplotlib. It is designed for statistical plotting.

c.3.2) **Correlation plot**

Correlation states how the features are related to each other or the target variable. Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable). Because we have many columns, a good way to quickly check correlations among columns is by visualizing the correlation matrix as a heatmap. Heatmap makes it easy to identify which features are most related to the target variable. We have plotted heatmap of correlated features using the seaborn library. From this distribution, we can derive the fact as follows :

- Employee Environment Satisfaction, Employee Last Salary High Percent and Worklife Balance are having the higher correlation when comparing to all features **.**They are positively correlated**. The increase of these features will increase Performance Rating**
- Employee Job Role, Years With Current Manager, Years Since Last Promotion, Experience Years In Current Role, Experience Years At this Company are negatively correlated : **The increase of these features will decrease Performance Rating**
- The coefficient correlation of Employee Environment Satisfaction, Employee Last Salary High Percent and Worklife Balance are respectively 0.4 (40 % in percentage), 0.33 (33 % in percentage), and 0.12 (12 % in percentage),
- The coefficient correlation of Years Since Last Promotion is -0.17 (17 % in percentage)

An overview of Machine Learning Model

The machine learning models used in this project are :

- **K-Nearest Neighbors** : K Nearest Neighbors is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classify a data point based on how its neighbours are classified.
- **Support Vector Machine** : Support Vector Machine is a machine learning technique that can be used for both regression and classification problems. It constructs a hyperplane in multi-dimensional space to separate a dataset into different classes in the best possible way.
- **Naive Bayes Bernoulli** : A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task.The crux of the classifier is based on the Bayes theorem.
- **Random Forest with GridSearchCV** : A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap

Aggregation. We will use grid search cross validation method to determine the optimal values to be used for the hyperparameters of our model from a specified range of values.

- **XGBoost Classifier** : XGBoost is an open source library providing a high-performance implementation of gradient boosted decision trees. An underlying C++ codebase combined with a Python interface sitting on top makes for an extremely powerful yet easy to implement package.

The predicted and test data achieved the accuracy rate of :

- **K-Nearest Neighbors** : 82%
- **Support Vector Machine** : 77.5%
- **Naive Bayes Bernoulli** : 72.5%
- **Random Forest with GridSearchCV** : 92%
- **XGBoost Classifier** : 91%

We can conclude that **Random Forest with GridSearchCV** performs the best follow by **XGBoost Classifier**.

Summary

The objective of our project is to provide the results from our analysis and machine learning model.

a) **Department Wise Performance**

In Employee department features, we have 6 departments available. The performance analysis by department as follows :

- **Sales** : For most of the employees The Performance rating is 3. Male employees perform a little bit than females
- **Human Resources** : For the maximum number of employees, the Performance Rating is 3. The females employees perform well in Human Resources department.
- **Development :** The majority of employees are under level 3 performance. Male and female employees have the same level of performance.

- **Data science :** Level 3 performance is very high here. Male employees perform also well.
- **Research & Development** : We can observe that different employees with distinct age are in every level of performance.
- **Finance** : With the increase of age, the finance employee performance is decreasing.

b) **Top 3 Important factors affecting employee performance**

By using **Random Forest Classifier** we can derive the importance of each variable on the tree decision. Ensembles of decision tree methods like **gradient boosting** can automatically provide estimates of feature importance from a trained predictive model. Top 3 Important factors affecting employee performance are :

- Employee Salary Hike Percentage
- Employment Environment Satisfaction
- Years Since The Last Promotion

c) **A trained model which can predict the employee performance based on factors as inputs**

There are two models which can predict the employee performance with the highest accuracy :

- Random Forest with GridSearchCV : 92%
- XGboost Classifier : 91%

d) **Recommendations to improve the employee performance based on insights from analysis.**

- The company should increase the salary of the employees
- The company should provide a more friendly, comfortable environment for the employees
- The company should help them to maintain a worklife balance

- Employees promotion must be planed by the company : It will help them to achieve more performance