**Employee Performance Analysis**

**INX Future Inc.**

* Candidate Name : Mihirkumar Harshadbhai Prajapati
* Candidate E-Mail : prajapatimihir502@gmail.com
* REP Name : DataMites™ Solutions Pvt Ltd
* Assesment ID : E10901-PR2-V18
* Module : Certified Data Scientist - Project
* Exam Format : Open Project- IABAC™ Project Submission
* Project Assessment : IABAC™
* Registered Trainer : Ashok Kumar A
* Submission Deadline Date: 08-April-2024

# PROJECT SUMMARY:

## BUISNESSCASE & GOAL OF PROJECT: BASED ON GIVEN FEATURE OF DATASET WE NEED TO PREDICT THE PERFOMANCE RATING OF EMPLOYEE

### INX Future Inc Employee Performance - Project

The Data science project which is given here is an analysis of employee performance.

**The Goal and Insights of the project are as follows:**

* Department wise performances
* Top 3 Important Factors effecting employee performance
* A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees
* Recommendations to improve the employee performance based on insights from analysis

The given Employee dataset consist of 1200 rows. The features present in the data are 28 columns. The shape of the dataset is 1200x28. The 28 features are classified into quantitative and qualitative where 19 features are quantitative (11 columns consists numeric data & 8 columns consists ordinal data) and 8 features are qualitative. EmpNumber consist alphanumerical data (distinct values) which doesn't play a role as a relevant feature for performance rating.

From Correlation we can get the important aspects of the data, Correlation between features and Performance Rating. Correlation is a statistical measure that expresses the extent to which two variables are linearly related. The analysis of the project has gone through the stage of Univariate, Bivariate & Multivariate analysis, correlation analysis and analysis by each department to satisfy the project goal.

The dataset consists of Categorical data and Numerical data. The Target variable consist of ordinal data, so this is a classification problem.The multiple machine learning model used in this project is Support vector classifier, Random forest classifier & Logistic Regression. from above all models Random forest classifier predicts higher accuracy 99.58%.

One of the important goal of this project is to find the important feature affecting the performance rating. The important features were predicted using the machine learning model feature importance technique. The main technique used in the preprocessing data using the Mannual & Frequency encoding method to convert the string - categorical data into numerical data, because, Most of machine learning methods are based on numerical methods where strings are not supportive. The overall project was performed and achieved the goals by using the machine learning model and visualization techniques.

### Data Preprocessing

Various preprocessing techniques were applied:

* **Missing value handling:** No missing values were present.
* **Outlier handling:** Detected and capped outliers in features like Total Work Experience In Years, Training Times Last Year etc using IQR ranges.
* **Feature scaling:** Scaled numerical features using StandardScaler to normalize distribution.
* **Label encoding:** Label encoding was used in the model to convert categorical variables into numerical format for analysis or modeling purposes.

### Exploratory Data Analysis

Performed detailed exploratory analysis:

* **Univariate analysis** on continuous features like Age, EmpHourlyRate using histograms. Analyzed categorical features like Department using bar plots.
* **Bivariate analysis** using correlation matrix and scatter plots between features like Age and EmpLastSalaryHikePercent.
* **Multivariate analysis** by generating pairplots to understand interactions between multiple features.

Key observations:

* Features like TotalWorkExperienceInYears exhibited right skewed distributions. Applied transformations like square root.
* Attributes like YearsSinceLastPromotion provided insights into career growth.

### Model Building and Evaluation

Experimented with 3 supervised ML models:

* **Support Vector Machine:** Achieved 97% testing accuracy without tuning using GridSearch.but I have provide GridSearch hyper parameters tuning.
* **Random Forest:** Random forest very well perform in training data with 100% accuracy but in testing 99.58%.
* **Logistic Regression:** Logistic Regression achieved 91.80 testing accuracy.
* Selected as **Random Forest** optimal model based on superior test performance.

### Model Diagnostics

* Calculated evaluation metrics like accuracy, precision, recall, F1-score. Generated classification reports.
* Created confusion matrix to analyze types of correct and incorrect predictions.
* Visualized feature distributions and correlations to diagnose data issues.

**Department wise performances**

* The exploratory analysis included bar plots visualizing employee distribution across different departments like Sales, R&D etc.
* This provided insights into department-wise composition and potential imbalances.

**Top 6 Important Factors effecting employee performance**

* 1.EmpLastSalaryHikePercent: Represents the percentage increase in an employee's salary during their last hike.
* 2.EmpEnvironmentSatisfaction: Reflects an employee's satisfaction level with their work environment.
* 3.YearsSinceLastPromotion: Measures the time since an employee was last promoted.
* 4.EmpWorkLifeBalance: Indicates the balance between an employee's work and personal life.
* 5.ExperienceYearsInCurrentRole: Represents the number of years an employee has spent in their current role.
* 6.EmpJobRole: Specifies the employee's specific job role or position within the organization.

**A trained model which can predict the employee performance**

* Three models (SVM, Random Forest and Logistic Regression) were trained and evaluated using accuracy metrics.
* The Random Forest was selected as the best performing model for employee performance prediction.

**Recommendations to improve the employee performance**

* This aspect was not covered in the report. Some example recommendations could be:
  + Increase training/upskilling programs to improve competencies
  + Foster better work relationships and engagement through team building
  + Offer monetary and non-monetary incentives linked to performance
  + Provide coaching and mentoring for low performers

### Conclusion

**The Random Forest model gave 99.58% test accuracy with good generalization capability. Followed a structured machine learning workflow involving data preprocessing, model building, diagnostics and optimizations. The end-to-end implementation, analysis and choice of final model were appropriate**.