Employee Performance Analysis

INX Future Inc.

- Candidate Name : Raja Prabu Manivel
- Candidate E-Mail : rpthalaprabhu@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 : 24-MAY-2025

PROJECT SUMMARY

This project focuses on analyzing employee performance data using supervised machine learning algorithms after thorough preprocessing and exploratory data analysis. The objective is to predict employee performance based on various features such as experience, education, job involvement, and other HR-related metrics.

---

1. Algorithms and Training Methods Used

The following classification algorithms were implemented and trained on the dataset:

> **Logistic Regression**
> A baseline linear classifier used to model the probability of class membership. Training involved splitting the data into training and test sets and applying standard scaling before fitting the model.
>
> **Random Forest Classifier**
> An ensemble learning method using multiple decision trees. This model was trained with the default RandomForestClassifier from scikit-learn and used to improve prediction performance through bagging and feature importance evaluation.
>
> **Support Vector Classifier (SVC)**
> A powerful kernel-based algorithm, trained with standard scaling to ensure that features were normalized. It's particularly useful in high-dimensional spaces and was used to test non-linear boundaries.

All models used a **train/test split**, typically 80/20, and were evaluated using **accuracy**, **classification reports**, and **confusion matrices**.

---

2. Most Important Features Selected and Why

No explicit feature reduction technique like **PCA** (Principal Component Analysis) or **Factor Analysis** was used. However, several features were engineered or selected based on domain knowledge and EDA findings:

> **JobInvolvement**, **PerformanceRating**, **YearsAtCompany**, and **Age** were identified as **highly correlated with the target variable**.
>
> **Categorical features** such as BusinessTravel, Education, Gender, Department, etc., were **one-hot encoded** or label encoded to convert them into numerical format.
>
> **Missing value handling** and **data scaling (StandardScaler)** were used to ensure clean and normalized data input for models.

These features were selected based on their **correlation with the target** and **business relevance** to performance evaluation.

---

3. Other Techniques and Tools Used

- **Data Preprocessing**: Handled null values, categorical encoding, and feature scaling.
- **Exploratory Data Analysis (EDA)**: Used seaborn and matplotlib for visualization of feature distributions, correlation heatmaps, and class imbalances.
- **Model Evaluation Tools**:
  - sklearn.metrics for classification reports and confusion matrices.
  - Accuracy score as a primary metric.
- **Python Libraries**:
  - pandas, numpy – data manipulation
  - matplotlib, seaborn – visualization
  - scikit-learn – ML modeling and preprocessing

FEATURE SELECTION / ENGINEERING

1. What were the most important features selected for analysis and why?

From exploratory data analysis and preprocessing:

- **JobInvolvement**: Positively correlated with performance rating, reflecting motivation and commitment.
- **YearsAtCompany** and **YearsInCurrentRole**: Indicate loyalty and experience in the company, often tied to performance.
- **Age**: Older employees tended to show more stability and higher performance in some cases.
- **OverTime**: Employees working overtime were observed to have higher or lower performance depending on job satisfaction.
- **JobSatisfaction** and **EnvironmentSatisfaction**: Directly influence morale and thus performance.

These features were prioritized based on:

Visual correlations in heatmaps.

Domain relevance (i.e., HR insights).

Model performance improvement when included.

---

2. Did you make any important feature transformations?

Yes, multiple transformations were done:

Label Encoding for binary categories like Gender, OverTime.

One-Hot Encoding for multi-class categories like Department, BusinessTravel, JobRole.

Standard Scaling using StandardScaler for numerical features (e.g., Age, YearsAtCompany) to prepare for SVM and Logistic Regression.

Handling Imbalanced Classes was partially addressed by checking value counts, although no SMOTE or resampling techniques were applied.

---

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

A **correlation matrix** was used to identify features like JobInvolvement, JobSatisfaction, and YearsAtCompany as positively associated with performance.

Multicollinearity wasn't deeply addressed (e.g., no VIF calculation), but interactions such as between OverTime and JobSatisfaction were explored in EDA plots.

Redundant or weakly correlated features like EmployeeNumber, EmployeeCount were dropped.

RESULTS, ANALYSIS AND INSIGHTS

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

Yes:

- **OverTime vs JobSatisfaction**: Employees working overtime were often less satisfied, yet sometimes rated high in performance, suggesting over-reliance on committed employees.

- **YearsSinceLastPromotion** had a subtle inverse relationship with performance in some cases, suggesting stagnation reduces motivation.

---

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

The **RandomForestClassifier** provided:

- Highest accuracy among models tested.

- Ability to identify and rank feature importance, improving model transparency.

Other essential techniques:

- **EDA for feature insights**

- **Standardization for SVC and Logistic Regression**

- **Encoding of categorical data**

---

3. Provide clear answers to the business problems mentioned in the project on the basis of analysis.

- **Which employees are most likely to perform well?**
  Employees with high job involvement, moderate tenure, positive satisfaction metrics, and occasional overtime perform better.

- **Which features most influence performance?**
  JobInvolvement, JobSatisfaction, OverTime, and YearsAtCompany were top predictors.

- **How can HR optimize resource allocation?**
  Focus training and promotion opportunities on employees with high involvement but lower satisfaction scores to boost retention and performance.

4. More business insights you gain from the analysis.

Departments with consistent overtime and low satisfaction need review — a high turnover risk.

Long-tenure employees not recently promoted show reduced engagement.

Gender and marital status showed weak correlation to performance — HR decisions should avoid biases