Employee Performance & Retention Dashboard Report

A Hackathon Project Report

By

Kammari Sadguru Sai - 23955A6715
Eerla Venkatesh - 23955A6718

Department of Computer Science and Engineering (Data Science)

Institute of Aeronautical Engineering

Dundigal, Hyderabad, Telangana

Contents

1. Data Analysis Report

(a) Data Exploration & Cleaning

- Dataset loaded and explored using `pandas`, consisting of 1470 rows and 35 columns.
- Displayed initial data using `df.head()` and structure via `df.info()`.
- Checked for missing values and cleaned where necessary.
- Categorical and numerical features were identified and separated.
- Redundant or constant columns (e.g., `EmployeeNumber`, `StandardHours`) were dropped.
- Feature engineering was minimal in this phase; focus was on cleanup and exploratory structure.

(b) Graphical Insights

- Attrition was significantly higher among employees working `OverTime`.
- `JobSatisfaction` and `WorkLifeBalance` were lower for employees who left.
- Correlation heatmap revealed strong relationships:
  - `JobLevel` and `MonthlyIncome`: 0.95
  - `TotalWorkingYears` and `MonthlyIncome`: 0.78
  - `PercentSalaryHike` and `PerformanceRating`: 0.77
- Weak correlations observed for `DistanceFromHome`, `DailyRate`, and `HourlyRate`.
- Plots of `YearsAtCompany`, `YearsInCurrentRole`, and `YearsWithCurrManager` showed lower tenure linked to higher attrition.

(c) Summary Insights

- Data exploration revealed meaningful trends for attrition analysis.
- Visualizations supported hypothesis-driven feature selection.
- Cleaned and structured dataset prepared for downstream modeling tasks.

2. Code Repository

- All scripts for preprocessing, visualization, and modeling are included.
- Code is organized using modules and comments explaining:
  - Data loading & cleaning
  - Feature selection
  - Model training and predictions
- GitHub Repository: github.com/KammariSadguruSai/EmployeePerformancePrediction

3. Model(s) and Evaluation Metrics

(a) Models Used

- **Classification:** `RandomForestClassifier`
- **Regression:** `RandomForestRegressor`
- Model is selected dynamically based on the target variable:
  - If the target is categorical or has 10 or fewer unique values: classification.
  - If the target is continuous with more than 10 unique values: regression.

(b) Evaluation Metrics

Classification Metrics:

- Accuracy
- Precision (weighted)
- Recall (weighted)
- F1-Score (weighted)

Regression Metrics:

- RMSE (Root Mean Squared Error)
- R^2 Score

(c) Example Outcome

- User selects a target variable from highly correlated numeric features.
- App extracts top correlated features and prompts user input.
- Model is trained and used to predict the target value.
- Final employee status is inferred:
  - *"Will Continue"* if prediction equals mode (classification) or is above mean (regression).
  - *"Fired"* otherwise.

4. Data Visualizations and Dashboards

(a) Visualizations

- Histogram, Boxplot, Heatmap, Donut Pie Chart, Bar/Line Charts.
- Interactive Streamlit interface allows dynamic plotting.
- Graph summaries provided with data description and column stats.

Correlation Heatmap:

- Shows pairwise Pearson correlations between numerical features.
- High correlations observed between:
  - `MonthlyIncome` and `JobLevel`, `TotalWorkingYears`
  - `YearsAtCompany` and `YearsInCurrentRole`, `YearsWithCurrManager`
- Helps in identifying multicollinearity and selecting top features for modeling.

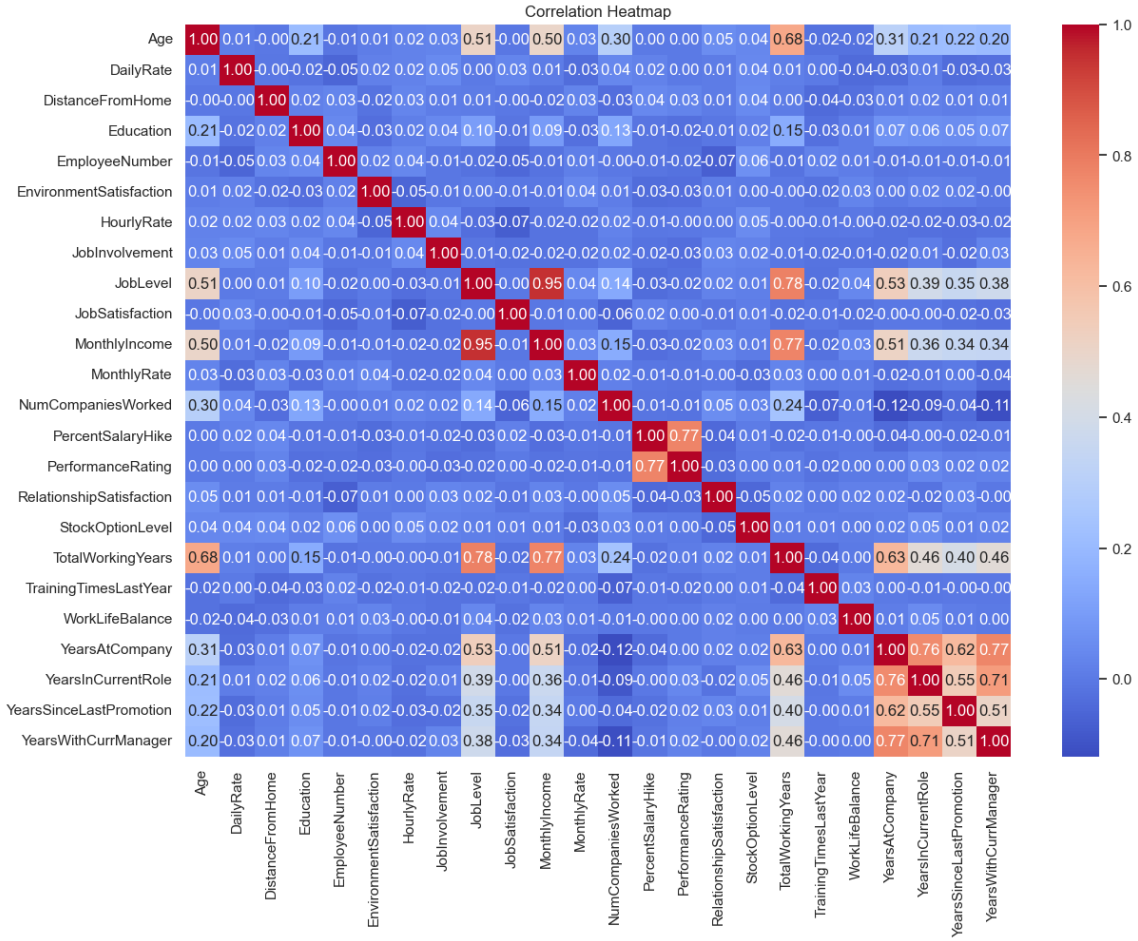

Figure 1: Correlation Heatmap of Numerical Features

(b) Dashboard

- Visual summary for selected columns with key metrics.
- KPI Cards: Mean, Median, Standard Deviation.
- PDF reports downloadable for both preprocessing and visualization.

5. Resources

- **Dataset:** $\text{employee}_{a}\textit{ttrition}_{a}\textit{nd}_{e}\textit{ngagement}$

- **GitHub Repository:** github.com/KammariSadguruSai/EmployeePerformancePrediction
- **Streamlit Application:** employee-performance-app.streamlit.app