Predictive Analytics for Employee Attrition

Authors: Nirja Rajeev, Kartik Badkas, Satheesh MK, Sarvesh Kulkarni, Saket Pitale

Abstract: Employee attrition poses significant challenges for organizations, affecting workforce stability and increasing recruitment costs. This study presents a predictive analytics framework to anticipate employee attrition using machine learning models. A publicly available HR dataset was preprocessed and evaluated using Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) classifiers. Key performance metrics such as accuracy, precision, recall, and ROC-AUC scores were used for evaluation. Feature importance analysis was conducted to identify critical variables influencing attrition. The proposed approach achieved high predictive performance and offers actionable insights for human resource management.

Keywords: Attrition Prediction, Machine Learning, Human Resources, Predictive Analytics, Employee Turnover, Classification Models

1. Introduction

Employee turnover is one of the critical challenges faced by organizations. Proactively identifying potential attrition cases allows HR departments to design retention strategies that improve employee satisfaction and organizational performance. This study explores predictive analytics to gain insights into factors influencing attrition and develop robust prediction models.

2. Literature Review

Previous studies have employed various techniques, including logistic regression, decision trees, and ensemble models, to predict attrition. IBM's HR Analytics Employee Attrition & Performance dataset has been a benchmark in several experiments. Research by Sharma et al. (2021) emphasized the role of OverTime and MonthlyIncome as key features.

3. Problem Statement

The goal is to build a predictive model that classifies whether an employee is likely to leave the organization, based on historical HR data.

4. Objectives

- To explore the dataset and understand attrition trends

- To preprocess and prepare data for machine learning models

- To train and evaluate multiple models for predicting attrition

- To identify key features affecting employee attrition

- To visualize insights and model performance

5. Methodology

This section includes steps from data preprocessing to model implementation.

5.1 Dataset

The dataset contains 1470 records and 35 features, including demographic, educational, professional, and compensation-related data.

Target Variable: Attrition (Yes/No)

5.2 Preprocessing

Steps include:

- Removal of irrelevant fields (EmployeeNumber, Over18, etc.)

- Label encoding of categorical features

- Scaling of features using StandardScaler

- Train-test split using 80:20 ratio

5.3 Model Implementation

Models used:

- Logistic Regression

```
# Fit logistic regression model
log_model = LogisticRegression(max_iter=2000)
log_model.fit(X_train, y_train)

# Predictions
y_pred = log_model.predict(X_test)
y_prob = log_model.predict_proba(X_test)[:, 1]
```

- Decision Tree

```
# Train Decision Tree model
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)

# Predictions
y_pred_dt = dt_model.predict(X_test)

# Evaluate model
accuracy_dt = accuracy_score(y_test, y_pred_dt)
report_dt = classification_report(y_test, y_pred_dt)

accuracy_dt, report_dt

print(accuracy_dt)
print(report_dt)
```

- Random Forest

```
# Train Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predictions
y_pred = rf_model.predict(X_test)
```

- Support Vector Machine (SVM)

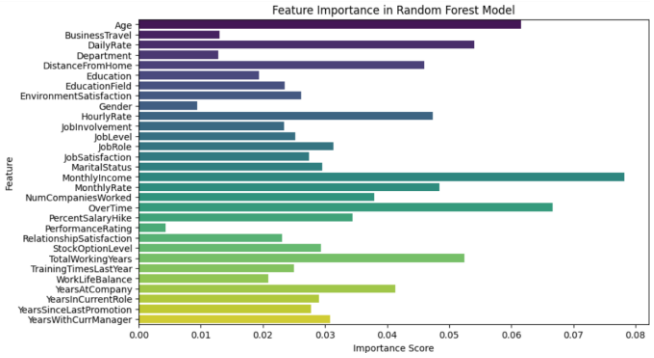
```
# Train SVM model
svm_model = SVC(kernel="rbf", random_state=42)
svm_model.fit(X_train, y_train)

# Predictions
y_pred = svm_model.predict(X_test)
```

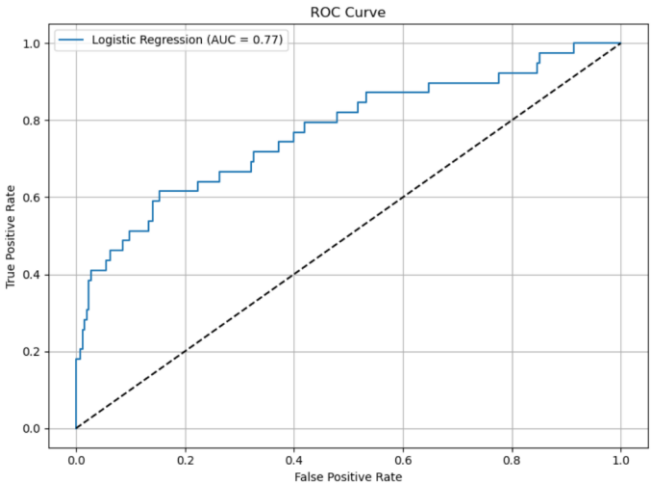
5.4 Evaluation Metrics

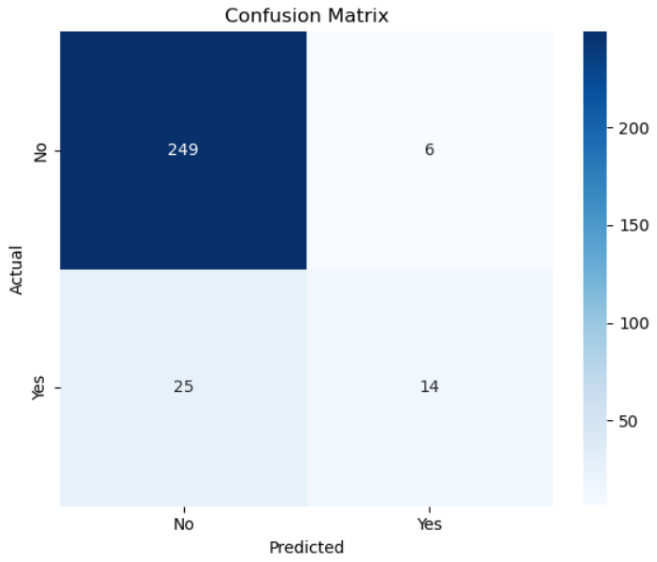
Metrics include Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix.

6. Results

Random Forest and SVM models achieved the highest prediction performance. Important features included OverTime, Age, and MonthlyIncome.







7. Discussion

Predictive models demonstrated strong capability in identifying high-risk employees. These insights can guide HR strategies and reduce turnover.

8. Conclusion

The machine learning models were effective in predicting employee attrition and identifying key influencing factors. This helps organizations in retaining talent.

9. Future Work

- Incorporate sentiment analysis from employee surveys

- Use deep learning or ensemble stacking

- Integrate predictions in real-time HR dashboards

10. References

Sharma, A., & Gupta, V. (2021). "Machine Learning Approaches for Employee Attrition Prediction: A Case Study." International Journal of Computer Applications

IBM HR Analytics Dataset – Kaggle

M. Dey, "A Comparative Study of Predictive Models for Employee Attrition," IEEE Access, 2020.