Employee Attrition Prediction and Analysis

Project Overview

This project analyzes and predicts employee attrition using the IBM HR Analytics dataset. Our goal was to identify key factors that contribute to employee attrition and build a predictive model that can assist HR departments in retaining valuable employees.

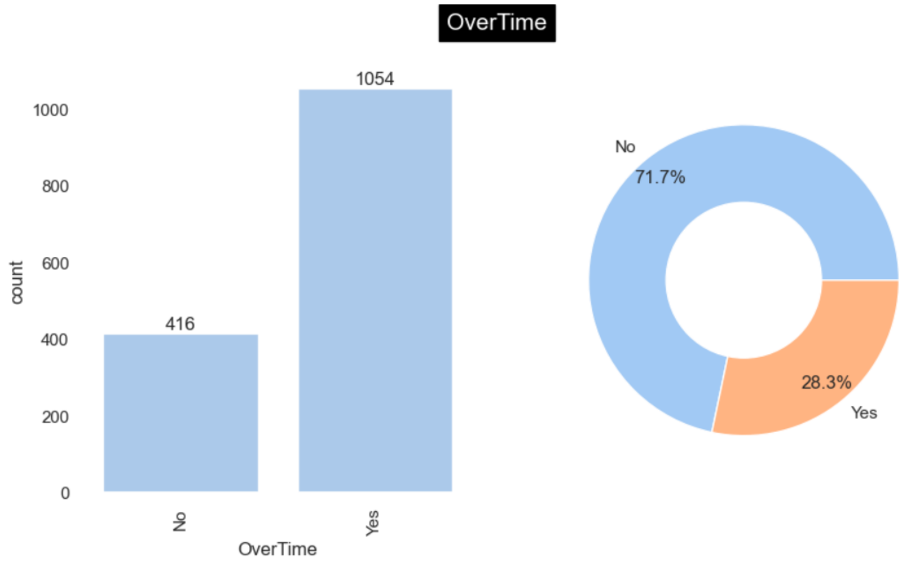
The project covers the entire data science pipeline, including data preprocessing, feature engineering, exploratory data analysis (EDA), handling class imbalance with SMOTE, training multiple machine learning models, and deploying the final model through a Streamlit application.

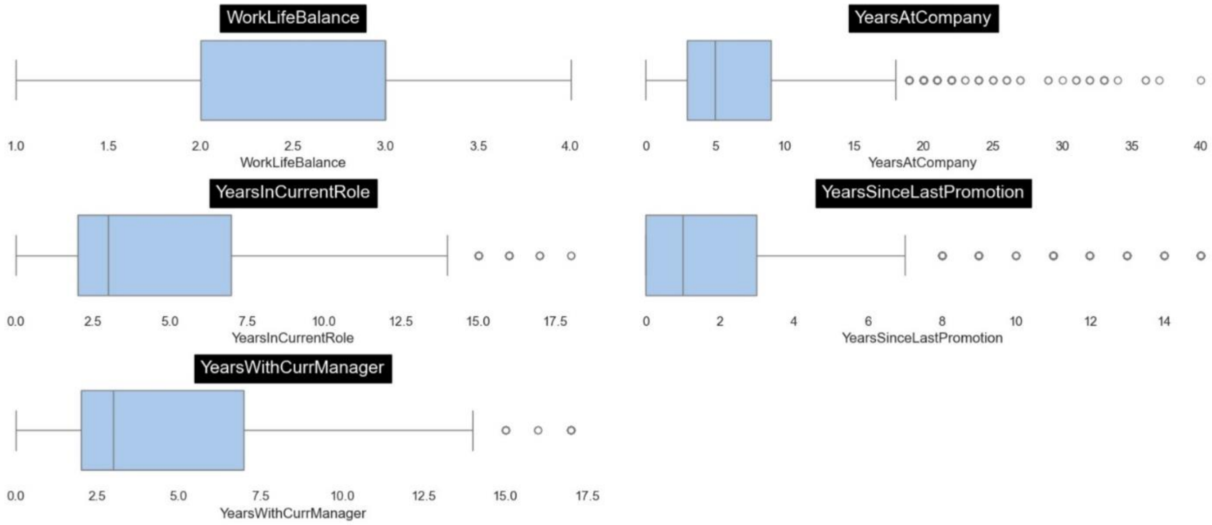
Dataset

- **Source:** Kaggle - IBM HR Analytics Attrition Dataset
- **Size:** 1,470 employees
- **Features:** 35 employee attributes, including Age, Gender, Job Role, Department, Overtime, Education, Monthly Income, Distance From Home, etc.

Exploratory Data Analysis (EDA)

- Countplots and pie charts were used to visualize categorical features.
- Boxplots were used to assess numerical feature distributions and detect outliers.





Key Observations

- Attrition is the highest for both men and women from 18 to 35 years of age and gradually decreases.
- As income increases, attrition decreases.
- Attrition is much, much less in divorced women.
- Attrition is higher for employees who usually travel than others, and this rate is higher for women than for men.
- Attrition is the highest for those in level 1 jobs.
- Women with the job position of manager, research director and technician laboratory have almost no attrition.
- Men with the position of sales expert have a lot of attrition.

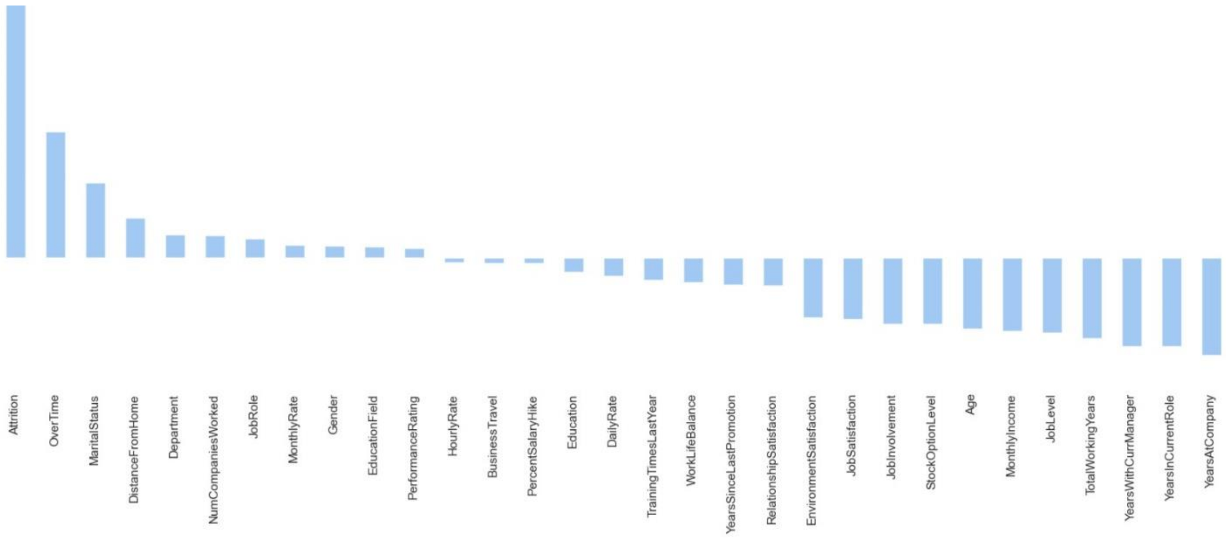
Data Preprocessing

- Removed columns with constant or irrelevant values: Over18, EmployeeNumber, EmployeeCount, and StandardHours.
- Handled outliers using **IQR Capping** for selected numerical features to avoid data loss: MonthlyIncome, TotalWorkingYears, etc
- Applied **Label Encoding** to categorical features

Correlation Analysis with Target Variable

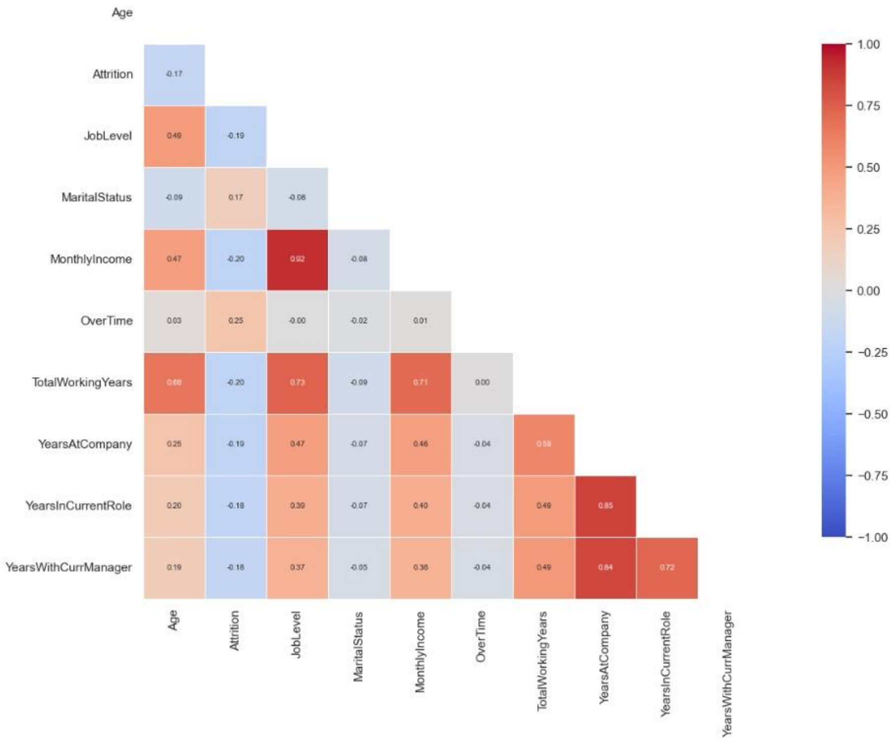
- After encoding, calculated **correlation** between all features and the target (Attrition) to identify significant predictors.

- Features with correlation **below 0.15** were **dropped** from the dataset to focus on stronger predictors.



Heatmap Visualization

- A **heatmap** was generated to visualize the final correlation matrix, showing relationships between selected features and the target.
- We **removed MonthlyIncome** and retained JobLevel as a more structured and informative feature.

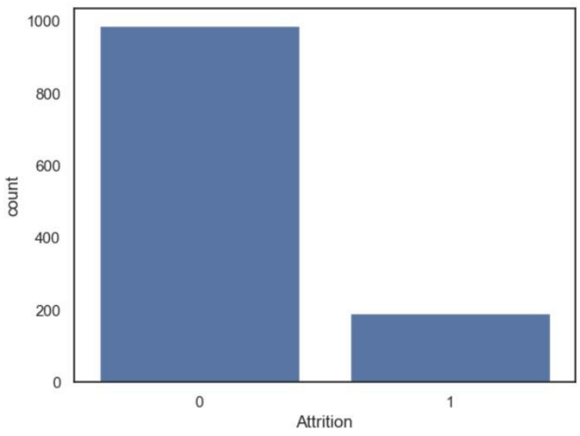


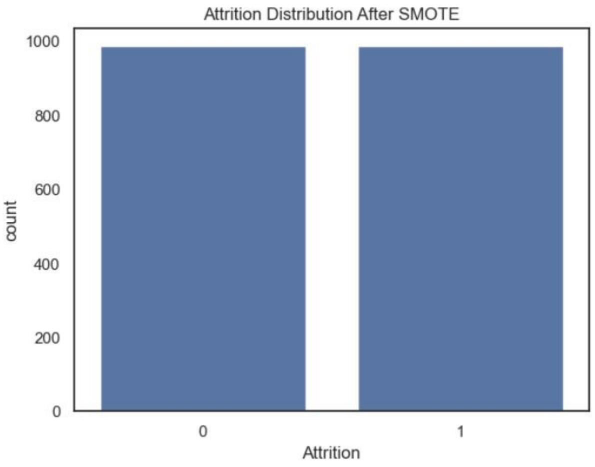
Feature Selection

- Choose 8 features with the highest correlation with the target Attrition:
  - Age
  - JobLevel
  - MaritalStatus
  - OverTime
  - TotalWorkingYears
  - YearsAtCompany
  - YearsInCurrentRole
  - YearsWithCurrManager

Handling Class Imbalance

- Used **SMOTE (Synthetic Minority Oversampling Technique)** to balance the dataset
- Before SMOTE: Attrition = No (84%), Yes (16%)
- After SMOTE: Balanced classes (50% each)



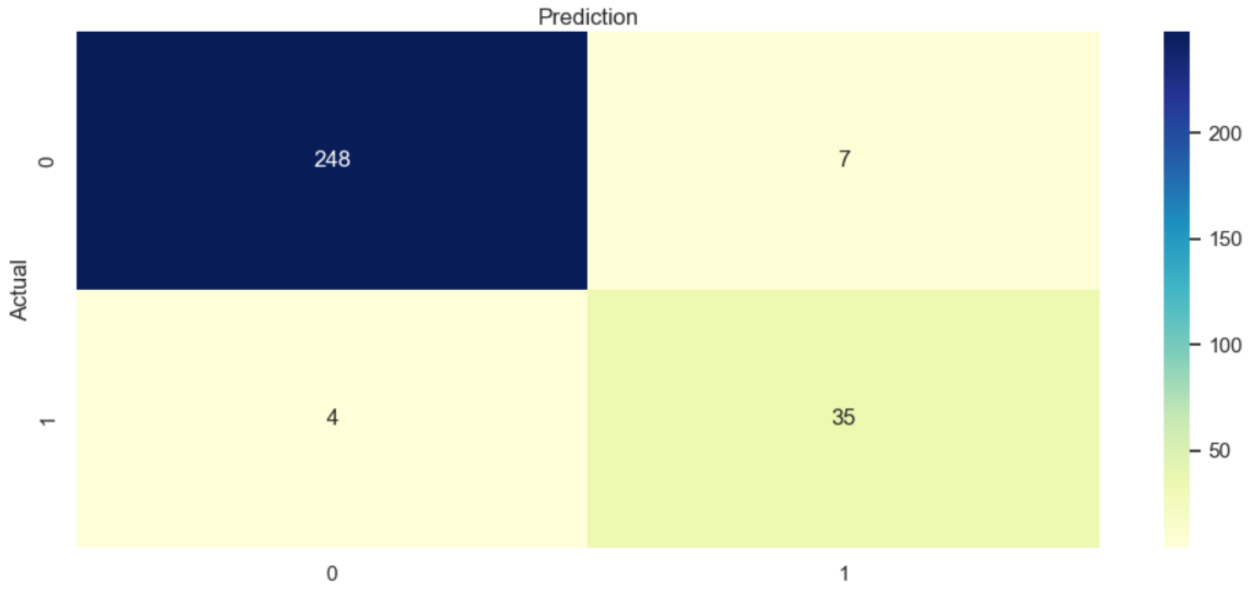

Attrition Distribution After SMOTE

Model Training & Evaluation

We trained and evaluated four models:

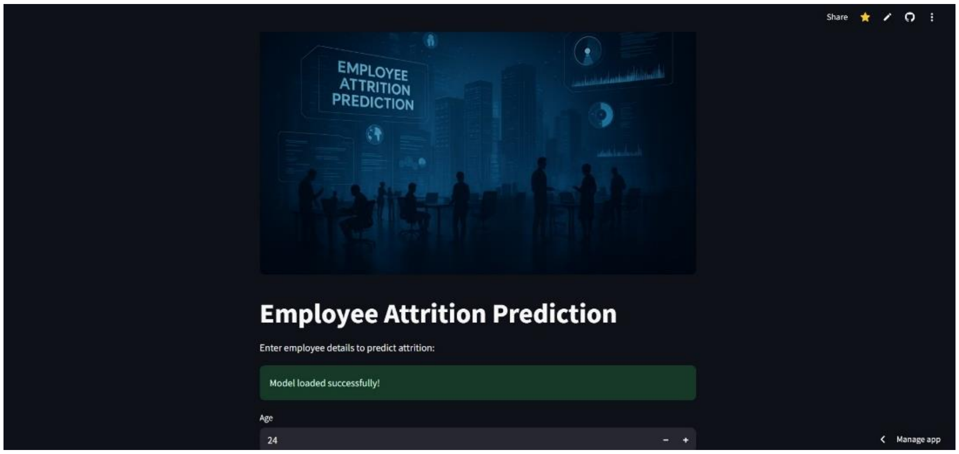
Model	Accuracy	F1 Score (Yes)	Recall	Precision
LogisticRegression	62.5%	0.42	0.80	0.28
DecisionTree	94.2%	0.81	0.88	0.72
GradientBoosting	95%	0.83	0.87	0.79
RandomForest (with GridSearchCV) #	96.2%	0.86	0.90	0.83

Confusion matrix:



Deployment

- Built an interactive **Streamlit application** to predict attrition risk
- Used only the selected 8 features as input
- Model and encoders were saved using joblib
- App supports real-time prediction based on user input
- Streamlit Link



Employee Attrition Prediction

Enter employee details to predict attrition:

Model loaded successfully!

Age

24

Challenges Faced

- Class imbalance caused biased predictions early on
- Dataset contained many features with weak correlation to target
- Presence of outliers skewed model training
- Small dataset size limited model complexity

Solutions

- Applied **SMOTE** to address class imbalance
- Selected 8 high-correlation features to simplify the model
- Used **IQR Capping** to reduce outlier impact without losing rows
- Chose **Random Forest** for its robustness with small datasets and non-normalized inputs

Conclusion

- The Random Forest model showed the best performance.
- Using SMOTE greatly improved results for minority class.
- Capping helped retain all data while managing outliers.
- Feature selection simplified the model while preserving accuracy.
- The Streamlit app offers a practical tool for HR to assess attrition risk interactively.

Recommendations

- Integrate the model with HR systems for automated prediction.
- Include satisfaction and engagement metrics in future versions.
- Monitor model drift and performance in production environment.