Forecasting Employee Attrition: Insights and Strategies

Capstone Two - Final Presentation
Springboard Data Science Career Track

Tuba Cakir Kranda

Problem Statement

- Employee attrition poses significant challenges for organizations
  - Impacts productivity, knowledge transfer, costs of hiring/training new employees
  - Disrupts team dynamics, morale and overall organizational performance
  - High attrition rates lead to loss of valuable human capital and expertise
- Objective: Analyze factors influencing attrition and develop a predictive model
  - Identify key factors influencing an employee's decision to leave
  - Develop a predictive model to forecast attrition risk for individual employees
  - Generate actionable insights to improve employee retention strategies
  - Enable proactive interventions to retain valuable talent

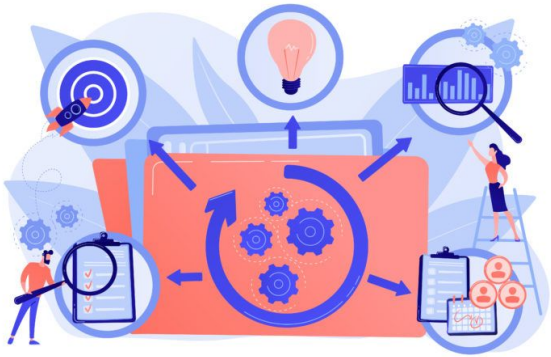
Dataset

- IBM HR Analytics Employee Attrition & Performance dataset (Kaggle)
- Contains 1470 employee records
- Features: age, education, job role, satisfaction levels, attrition status, etc.

```
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
```

Methodology

- **Data Wrangling**
  - Ensure data cleanliness and completeness.
- **Exploratory Data Analysis (EDA)**
  - Explore distributions, correlations, and trends.
- **Pre-processing**
  - Encode categorical variables and scale features.
- **Modeling**
  - Train and evaluate machine learning models.
- **Analysis**
  - Interpret model performance and feature importance.

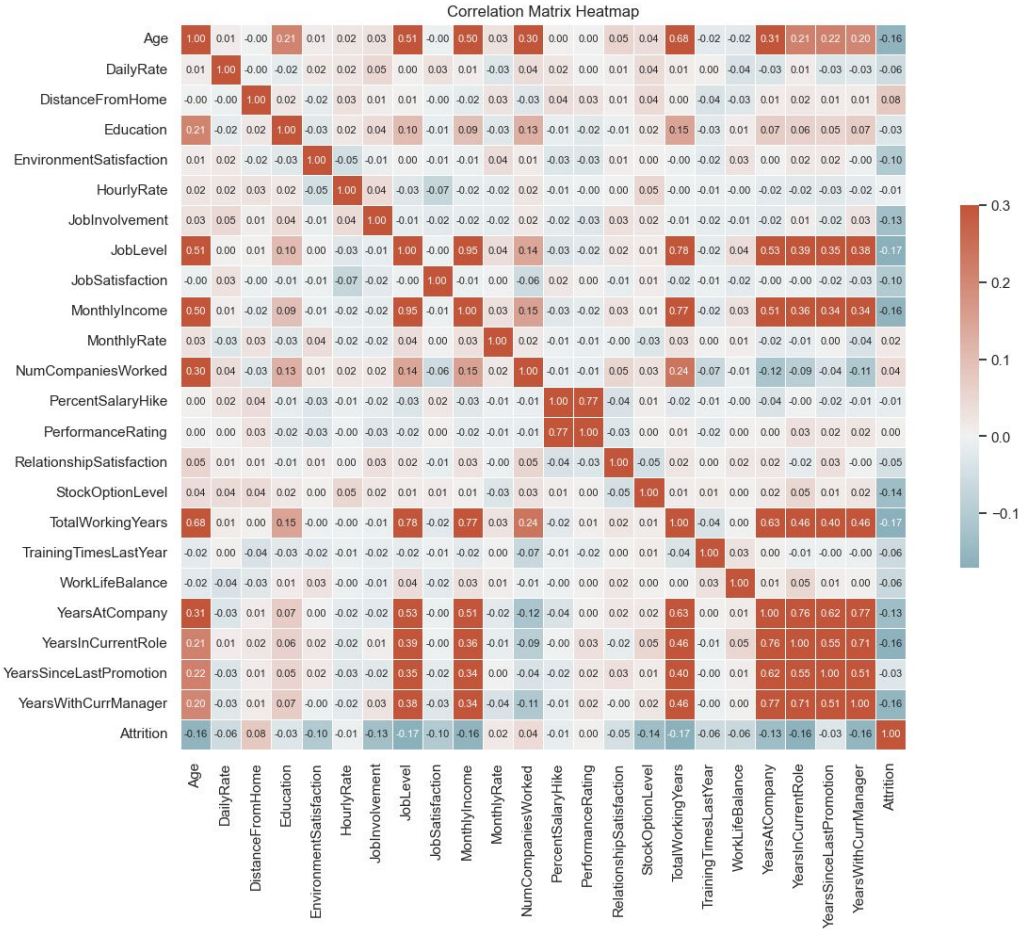


Data Wrangling

- Data Collection
- Data Organization
- Data Definition
  - Summary statistics for numeric columns to understand distributions.
  - Unique value counts for categorical columns to identify distinct categories.
  - Calculating percentages of unique values for categorical columns to understand distribution.
- Data Cleaning
  - Checking missing values to identify any gaps.
  - Visualizing missing data patterns.
  - Identifying duplicate rows to ensure data integrity.
  - Dropping columns with zero variance ('EmployeeCount', 'Over18', 'StandardHours').

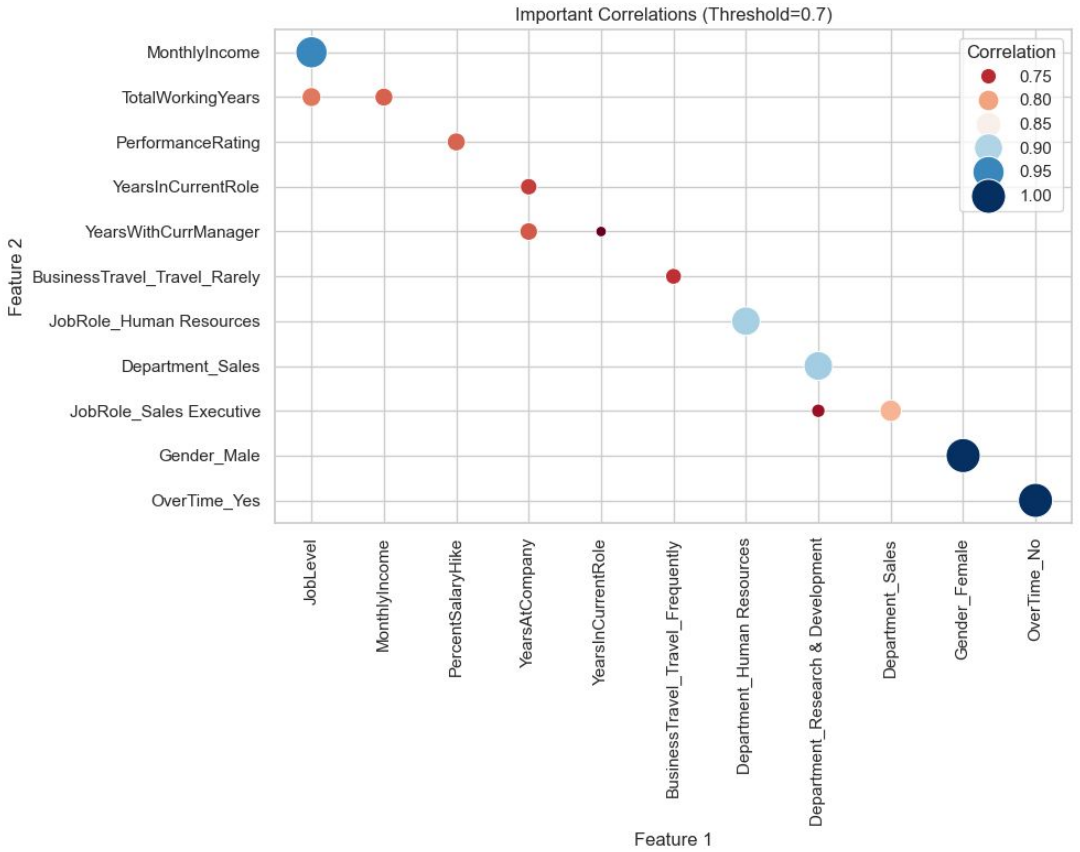
EDA Highlights

- Descriptive statistics
  - mean, median, std. dev.
- Visualizations
  - histograms, boxplots, etc.
- Correlation analysis
  - identified key influencers



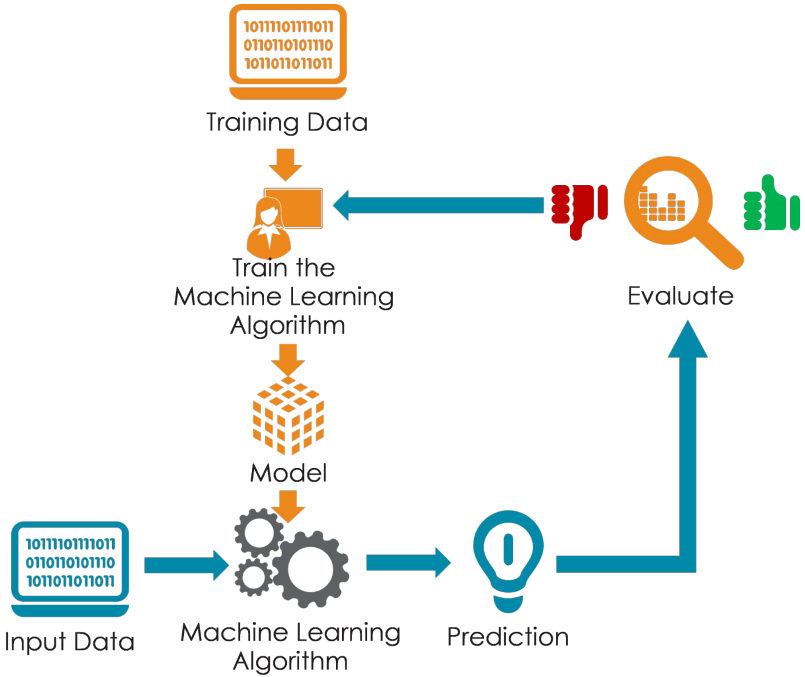
Data Pre-processing

- One-hot encoding for categorical variables
- Feature scaling with StandardScaler
- 75% training, 25% testing split



Modeling

- Models:
  - Logistic Regression,
  - K-Nearest Neighbors,
  - Support Vector Machine (SVM),
  - Random Forest,
  - Naive Bayes,
  - Gradient Boosting,
  - XGBoost
- Hyperparameter tuning
  - GridSearchCV
- Evaluation Metrics
  - Accuracy, F1-score,
  - Precision, Recall, ROC AUC



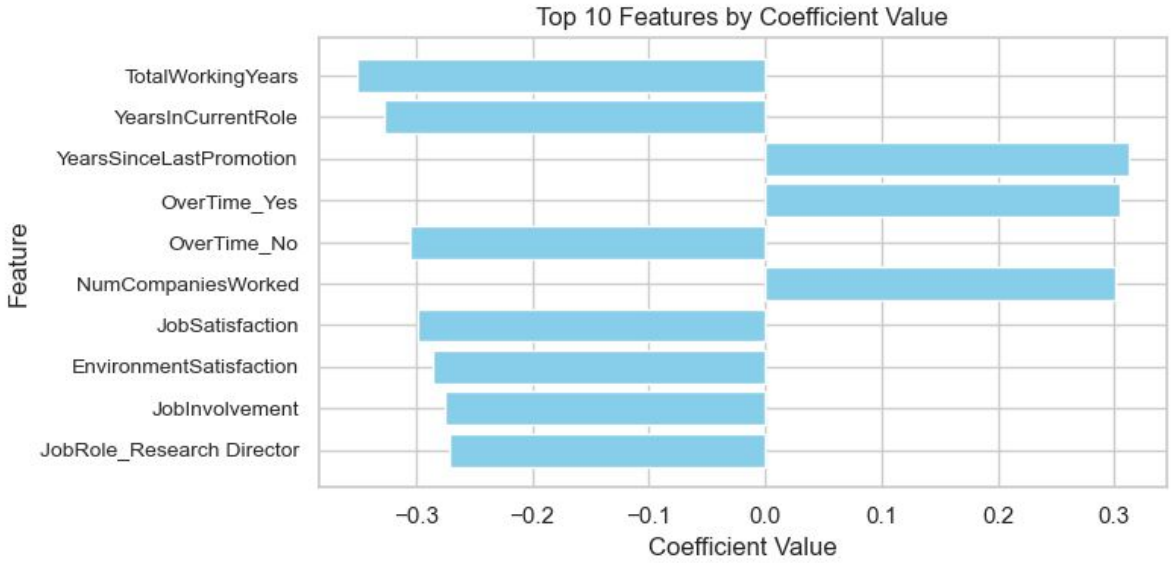
Best Performing Model

- Linear SVM (C=0.1)
- Accuracy: 89.9%
- F1-Score: 0.519

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
2	Support Vector Machine	0.899457	0.689655	0.416667	0.519481	0.694271
0	Logistic Regression	0.896739	0.678571	0.395833	0.500000	0.683854
1	K-Nearest Neighbor	0.891304	0.681818	0.312500	0.428571	0.645313
7	XGBoost	0.888587	0.684211	0.270833	0.388060	0.626042
6	Gradient Boosting	0.883152	0.592593	0.333333	0.426667	0.649479
3	Random Forest (Entropy)	0.869565	0.500000	0.083333	0.142857	0.535417
4	Random Forest (Gini)	0.869565	0.500000	0.083333	0.142857	0.535417
5	Naive Bayes	0.682065	0.228346	0.604167	0.331429	0.648958

Feature Importance

- SVM coefficients highlighting top predictive features

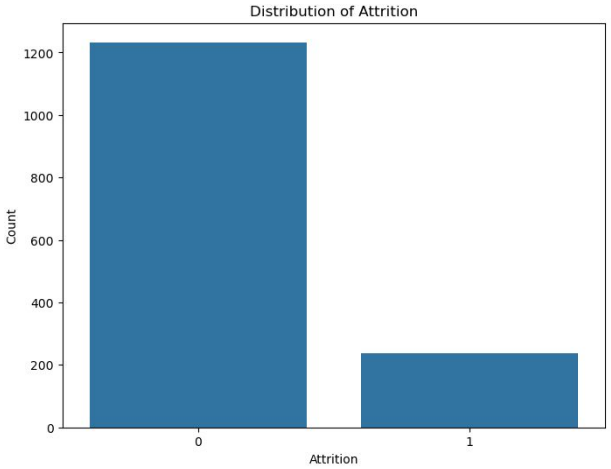


Key Findings

- **NumCompaniesWorked**, **YearsSinceLastPromotion**, and **OverTime_Yes**
  - relatively higher positive coefficients,
  - indicating their significant role in predicting attrition
- **TotalWorkingYears**, **YearsInCurrentRole**, **OverTime_No**, and **JobSatisfaction**
  - negative coefficients,
  - suggesting that lower levels of satisfaction and involvement are associated with higher attrition rates

Handling Class Imbalance

- Applied SMOTE (Synthetic Minority Over-sampling Technique)
- Increased recall for minority class (attrition)
- Tradeoff with decreased precision, more false positives



SVM (with SMOTE) Classification Report:

	precision	recall	f1-score	support
0	0.91	0.96	0.93	320
1	0.56	0.38	0.45	48
accuracy			0.88	368
macro avg	0.74	0.67	0.69	368
weighted avg	0.87	0.88	0.87	368

Recommendations

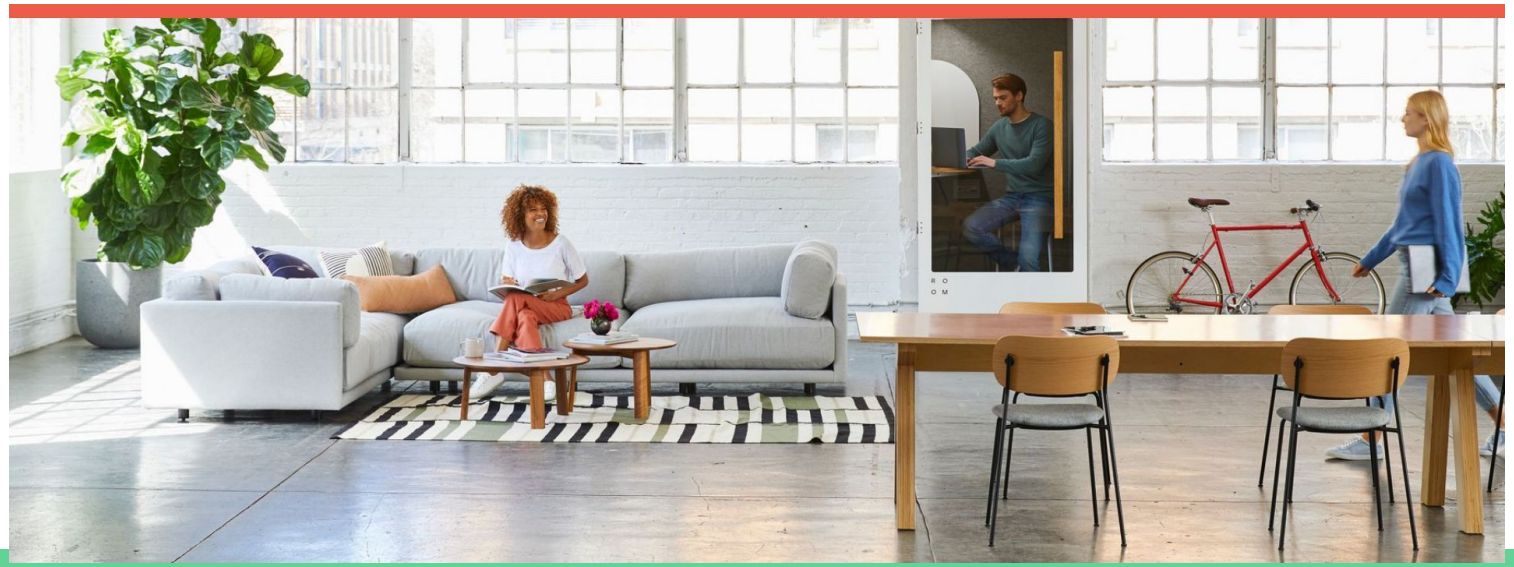
- Enhance work-life balance initiatives
- Develop career growth and mentorship programs
- Implement engagement surveys and feedback mechanisms
- Offer competitive compensation and benefits

Limitations & Future Work

- Potential gaps or biases in dataset
- Incorporate additional data sources (performance reviews, etc.)
- Explore advanced machine learning techniques

Conclusion

- Predictive model can identify at-risk employees proactively
- Actionable recommendations for retention strategies
- Optimization of HR processes and organizational success



References

- Dataset: Kaggle
- GitHub Repository: Link