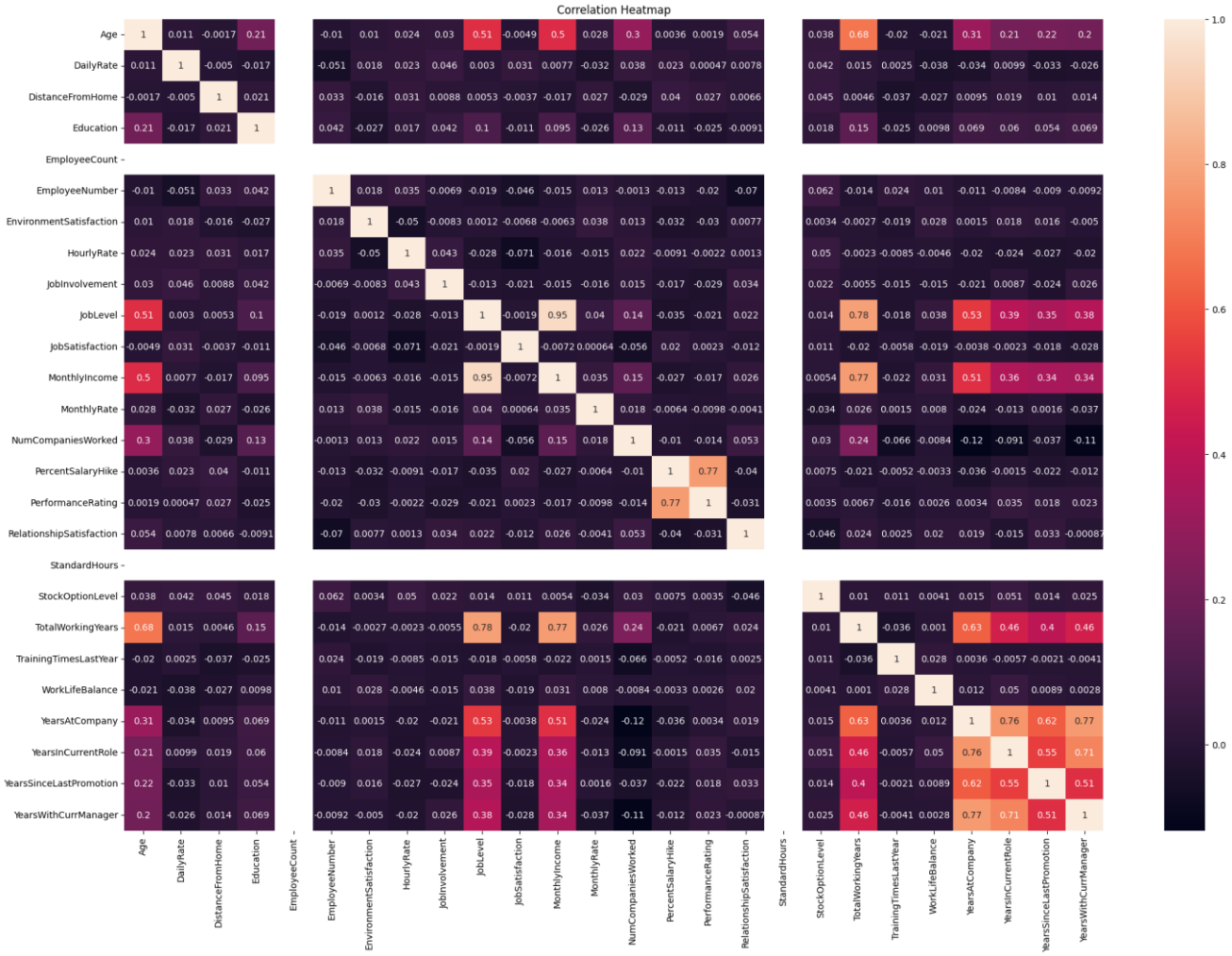
HR Employee Attrition Prediction Report

Introduction

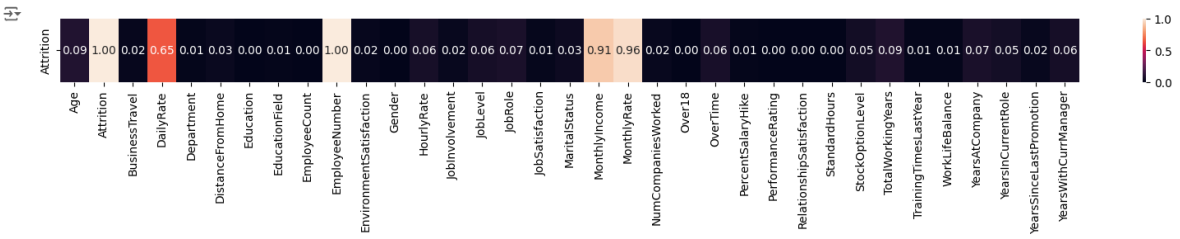
Employee attrition is a critical challenge faced by organizations across various industries. Understanding the factors contributing to attrition and predicting potential attrition cases can help companies take proactive measures to retain valuable employees. In this report, I will present an analysis of HR employee attrition data and develop predictive models to identify potential attrition cases.

Dataset Analysis

The dataset used in this analysis contains information about HR employee attrition, including demographic details, job role, salary, performance ratings, and other relevant attributes. Upon loading the dataset, we performed initial exploratory data analysis (EDA) to understand its structure, data types, and missing values. I also visualized the correlation matrix to identify relationships between features.



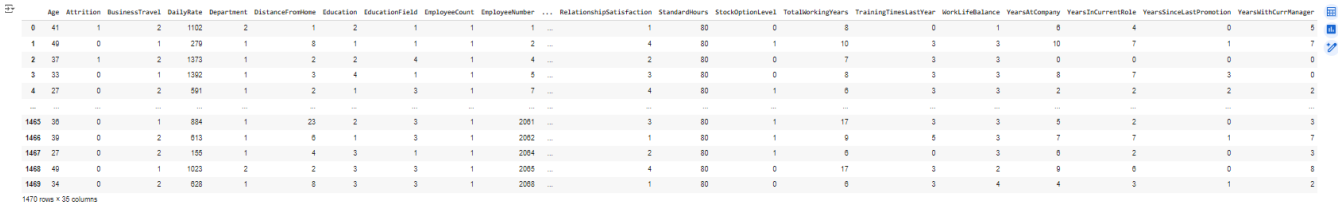
Co-relation between numeric features



Theilu's coefficient which gives association of each feature with Attrition column

Preprocessing Steps

Before building predictive models, we pre-processed the dataset by encoding categorical variables, handling missing values, and standardizing numerical features. We applied label encoding to categorical columns and filled missing values with zeros. Additionally, we standardized the numerical features using the StandardScaler to ensure uniform scaling.



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSatisfaction	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
0	41	1	2	1102	2	1	2	1	1	1	...	1	80	0	8	0	1	6	4	0	5
1	49	0	1	279	1	8	1	1	1	2	...	4	80	1	10	3	3	10	7	1	7
2	37	1	2	1373	1	2	2	4	1	4	...	2	80	0	7	3	3	0	0	0	0
3	33	0	1	1392	1	3	4	1	1	5	...	3	80	0	8	3	3	8	7	3	0
4	27	0	2	591	1	2	1	3	1	7	...	4	80	1	6	3	3	2	2	2	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1465	36	0	1	884	1	23	2	3	1	2061	...	3	80	1	17	3	3	5	2	0	3
1466	39	0	2	613	1	6	1	3	1	2062	...	1	80	1	9	5	3	7	7	1	7
1467	27	0	2	155	1	4	3	1	1	2064	...	2	80	1	6	0	3	6	2	0	3
1468	49	0	1	1023	2	2	3	3	1	2065	...	4	80	0	17	3	2	9	6	0	8
1469	34	0	2	628	1	8	3	3	1	2068	...	1	80	0	6	3	4	4	3	1	2

1470 rows × 35 columns

Dataset pre-processed

```
[49] Y_Train

array([0, 0, 1, ..., 1, 0, 0])

     X_Train

array([[ 1.41369115,  0.59277912,  0.79421172, ..., -0.05899761,
        -0.36030992, -0.28567748],
       [-0.09834647,  0.59277912, -1.44072151, ..., -0.60592139,
        -0.68214924, -0.85372023],
       [-1.71838678,  0.59277912, -1.14354907, ..., -1.15284518,
        -0.36030992, -1.13774161],
       ...,
       [-1.61038409, -0.92079337,  1.11348789, ..., -1.15284518,
        -0.68214924, -1.13774161],
       [-0.85436528,  0.59277912,  1.41311631, ..., -0.3324595 ,
        -0.68214924, -0.28567748],
       [ 1.41369115,  0.59277912, -1.32283492, ..., -1.15284518,
        -0.68214924, -1.13774161]])
```

Performed StandScaler for feature scaling

Model Development

I developed predictive models using logistic regression, logistic regression with L2 regularization, random forest classifier, and support vector machine (SVM) with RBF kernel. Later, I realized that the dataset is an imbalanced dataset with 237 belonging to one class and 1233 belonging to another class. For addressing class imbalance, I applied the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class.

Evaluation Results

Logistic Regression with L2 regularization:

- Accuracy: 0.8662131519274376

Random Forest Classifier (with SMOTE):

- Accuracy: 0.84

Support Vector Machine (RBF Kernel) (with SMOTE):

- Accuracy: 0.84

Optimization Techniques

To address class imbalance, we applied the SMOTE technique, which generated synthetic samples for the minority class to balance the class distribution. This helped improve the performance of the models, especially for the minority class.

Findings and Insights

- Logistic regression with L2 regularization achieved an accuracy of 0.87 in predicting employee attrition. This indicates that the model was able to correctly classify approximately 87% of the instances in the test dataset.
- While logistic regression is a simple yet effective algorithm for binary classification tasks, the performance could potentially be improved by exploring more complex models or refining feature engineering techniques.
- The random forest classifier, enhanced with the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, achieved an accuracy of 0.84.
- Despite its ability to handle non-linearity and interactions between features, the model struggled to accurately identify instances of employee attrition, particularly in the minority class.
- The support vector machine (SVM) with the radial basis function (RBF) kernel, combined with SMOTE for class balancing, also achieved an accuracy of 0.84.
- While SVMs are powerful classifiers, their performance can be sensitive to the choice of kernel and hyperparameters. Further tuning or exploring alternative kernel functions may be necessary to improve predictive performance.

Recommendations

- Based on my analysis, I recommend the following strategies to reduce employee attrition:
- Implement targeted retention programs based on key factors identified in the analysis.

- Improve communication and engagement with employees to address concerns and enhance job satisfaction.
- Provide opportunities for skill development and career advancement to promote employee growth and loyalty.
- Regularly monitor employee feedback and satisfaction metrics to identify potential attrition risks early.

Conclusion

In conclusion, the analysis highlights the importance of predictive modeling in identifying potential employee attrition cases. Employee attrition prediction is a challenging task due to the complex interplay of various factors influencing employee turnover, including job satisfaction, work-life balance, career growth opportunities, and organizational culture.

Class imbalance poses a significant challenge in developing accurate attrition prediction models, as minority class instances are often underrepresented in the dataset. Techniques such as SMOTE can help mitigate this issue by generating synthetic samples for the minority class.

While machine learning models can provide valuable insights into attrition risk factors, they should be used in conjunction with qualitative assessments and domain expertise to make informed decisions and develop effective retention strategies.