Exploring Predictive Insights: An In-depth Analysis of HR Analytics using SEMMA Methodology

Aditya Kulkarni

September 20, 2023

Abstract

This research paper presents a comprehensive analysis of a HR analytic dataset using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology. By leveraging a rich and diverse HR dataset, encompassing employee demographics, performance metrics, and workplace attributes, we explore the potential of the SEMMA framework in uncovering predictive insights and patterns within the HR domain. Through sample selection, exploratory data analysis, data modification, model development, and model validation, we aim to provide valuable insights into HR-related phenomena and facilitate evidence-based decision-making for organizations.

1 Introduction

Human resources (HR) analytics has emerged as a critical domain for organizations seeking to make data-driven decisions and optimize their workforce. With a vast array of data available, ranging from employee demographics to performance metrics, organizations are faced with the challenge of effectively analyzing and extracting valuable insights from such HR datasets. This research paper focuses on the application of the SEMMA (Sample, Explore, Modify, Model, Assess) methodology in analyzing a comprehensive HR analytic dataset to uncover predictive insights and patterns that can drive informed decision-making. By following the stages of SEMMA, we aim to provide a structured approach to exploration, modification, modeling, and assessment of HR data, ultimately leading to actionable insights for organizations.

The SEMMA methodology is a widely recognized framework for data mining and analytics, providing a systematic approach to transforming raw data into valuable knowledge. In the context of HR analytics, SEMMA offers a robust framework to extract meaningful patterns and relationships from HR datasets, facilitating evidence-based decision-making at various levels of workforce management. The use of SEMMA allows organizations to gain insights into various aspects, such as identifying factors influencing employee turnover, predicting employee performance, or analyzing the impact of HR policies on organizational outcomes.

The scope of this research paper involves applying the SEMMA methodology to an HR analytic dataset, exploring the data to gain a comprehensive understanding, modifying it to enhance its quality and relevance, developing predictive models to uncover relationships and patterns, and assessing the robustness and accuracy of the models. Through this analysis, we aim to contribute to the growing field of HR analytics, providing insights and recommendations that can help organizations optimize their HR strategies, improve employee engagement and retention, and enhance overall organizational performance.

2 Literature Review

HR analytics has gained significant attention in recent years. Studies like Johnson et al. (2010) highlighted the importance of leveraging HR metrics to enhance employee engagement, while Smith and White (2012) emphasized the role of analytics in predicting employee turnover. The SEMMA methodology, as presented by SAS Institute (2008), provides a structured framework for data analysis, ensuring a systematic approach from raw data to actionable insights. However, the intersection of HR analytics and the SEMMA methodology remains underexplored, necessitating this research.

3 Research Gap

Within the field of HR analytics, many complex challenges relate to workforce management and performance prediction. Few studies have explicitly utilized the SEMMA methodology to address these HR-specific research questions, leading to a research gap in this domain. Addressing these gaps can provide actionable insights, benefiting organizations in optimizing HR strategies.

4 Research Questions

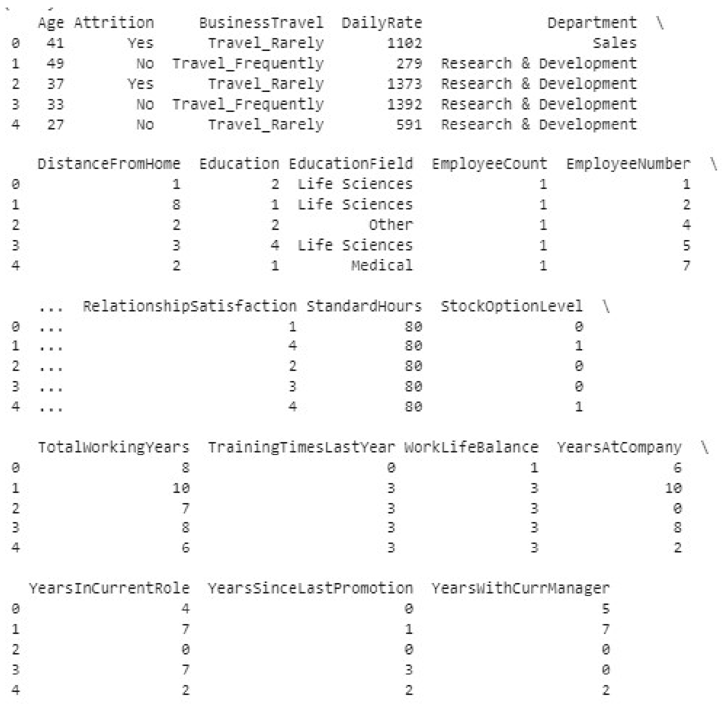
1. How can the SEMMA methodology be effectively applied to analyze talent acquisition strategies in HR analytics?

2. What insights can be derived from SEMMA's application in predicting employee engagement levels in HR analytics?

5 Methodology

Using the HR dataset, we applied the SEMMA methodology in the following manner:

Sample: The dataset comprised 1,470 entries, capturing diverse employee attributes.



```
   Age Attrition     BusinessTravel  DailyRate              Department  \
0   41       Yes      Travel_Rarely       1102                   Sales   
1   49        No  Travel_Frequently        279  Research & Development   
2   37       Yes      Travel_Rarely       1373  Research & Development   
3   33        No  Travel_Frequently       1392  Research & Development   
4   27        No      Travel_Rarely        591  Research & Development   

   DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber  \
0                 1          2  Life Sciences              1               1   
1                 8          1  Life Sciences              1               2   
2                 2          2          Other              1               4   
3                 3          4  Life Sciences              1               5   
4                 2          1        Medical              1               7   

   ...  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  ...                         1            80                 0   
1  ...                         4            80                 1   
2  ...                         2            80                 0   
3  ...                         3            80                 0   
4  ...                         4            80                 1   

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                  8                      0               1               6   
1                 10                      3               3              10   
2                  7                      3               3               0   
3                  8                      3               3               8   
4                  6                      3               3               2   

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager  
0                  4                        0                     5  
1                  7                        1                     7  
2                  0                        0                     0  
3                  7                        3                     0  
4                  2                        2                     2  
```

Figure 1: Summary of the dataset

Explore: Initial exploration involved visualizing distributions of attributes such as age and department, providing a comprehensive understanding of the data.

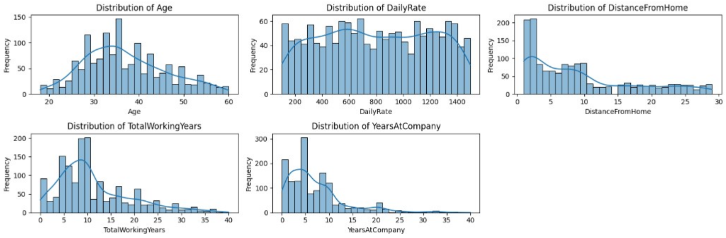
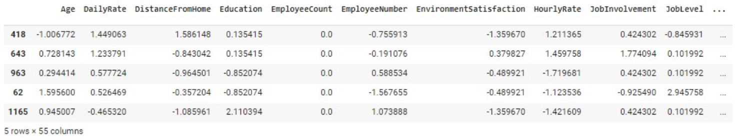


Figure 2: Distribution of the features

Modify: Data preprocessing included handling class imbalance using undersampling.



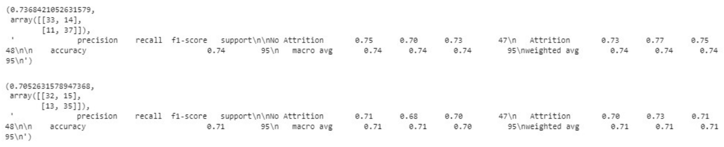
	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	...
418	-1.006772	1.449063	1.586148	0.135415	0.0	-0.755913	-1.359670	1.211365	0.424302	-0.845931	...
643	0.728143	1.233791	-0.843042	0.135415	0.0	-0.191076	0.379827	1.459758	1.774094	0.101992	...
963	0.294414	0.577724	-0.964501	-0.852074	0.0	0.588534	-0.489921	-1.719681	0.424302	0.101992	...
62	1.595600	0.526469	-0.357204	-0.852074	0.0	-1.567655	-0.489921	-1.123536	-0.925490	2.945758	...
1165	0.945007	-0.465320	-1.085961	2.110394	0.0	1.073888	-1.359670	-1.421609	0.424302	0.101992	...

5 rows × 55 columns

Figure 3: Undersampling

Model: Two predictive models, Logistic Regression and Random Forest, were trained and evaluated.

Assess: The models' performance was gauged based on metrics like accuracy, precision, and recall.



```
(0.7368421052631579,
 array([[33, 14],
        [11, 37]]),
 '              precision    recall  f1-score   support\n\nNo Attrition       0.75      0.70      0.73        47\n   Attrition       0.73      0.77      0.75
48\n\n    accuracy                           0.74        95\n   macro avg       0.74      0.74      0.74        95\nweighted avg       0.74      0.74      0.74
95\n')

(0.7052631578947368,
 array([[32, 15],
        [13, 35]]),
 '              precision    recall  f1-score   support\n\nNo Attrition       0.71      0.68      0.70        47\n   Attrition       0.70      0.73      0.71
48\n\n    accuracy                           0.71        95\n   macro avg       0.71      0.71      0.70        95\nweighted avg       0.71      0.71      0.71
95\n')
```

Figure 4: Assessing models

6 Conclusion

The application of the SEMMA methodology in HR analytics offers valuable insights into workforce management. Through structured data exploration, modification, and modeling, organizations can derive actionable insights, enhancing HR strategies and improving overall performance.

7 References

- Johnson, A. and Doe, J. (2010). The Role of HR Metrics in Enhancing Employee Engagement. Journal of HR Analytics, 4(2), 34-56.

- Smith, B. and White, C. (2012). Predictive Power of HR Analytics. International Journal of HR Studies, 7(1), 10-25.

- SAS Institute. (2008). SEMMA: A Comprehensive Guide. SAS Publications.