Analysis of HR Dataset to assess the Performance and Attrition of Employees

Group - 22

Mutasim Mahmud, 0813161
Applied Modelling & Quantitative Methods
(BigData)
Trent University
(Trail College)
Peterborough, Canada
mutasimmahmud@trentu.ca

Oluwatobiloba Ajibola, 0805333
Applied Modelling & Quantitative Methods
(BigData)
Trent University
(Trail College)
Peterborough, Canada
ajibolaoluatobiloba@trentu.ca

Muhammad Salman Hameed, 0777284
Applied Modelling & Quantitative Methods
(Big Data Financial Analytics)
Trent University
(Trail College)
Peterborough, Canada
muhammadhameed@trentu.ca

***Abstract*— The project aims to explore the analysis of an employee performance dataset sourced from Kaggle. The study will use Exploratory Data Analysis (EDA) and visualization techniques to infer useful insights, especially on attrition. It has become pertinent to understand and evaluate the performance of employees in a remote and hybrid work culture due to the thrust of Covid-19. Traditional methods of performance assessment have become either inefficient or ineffective; therefore, newer ideas are required.**

The dataset will first be pre-processed with Python to clean and make it ready for analysis. Next, Exploratory Data Analysis (EDA) will be carried out to study the patterns and relationships among data features by making use of visualization libraries like ggplot, seaborn, and matplotlib. It aims to communicate the findings clearly. Important features identified while performing EDA will be incorporated into the predictive model. For the employee attrition prediction model, we'll be making using of the XGBoost classifier algorithm. We will also be using several model evaluation techniques to check the performance of our model.

The expected outcomes of the project include a comprehensive understanding of the factors explaining employee performance, informative visualizations with clear depictions of the insights, and a predictive model capable of accurately assessing whether or not an employee will leave the company. In this direction, the current study attempts to contribute valuable information to human resource management in the provision of analytic insights and tools to facilitate improved processes of employee appraisal. Eventually, the output will help organizations upgrade their work processes to adapt to a transformed environment while being equally productive.

Keywords—machine learning, attrition, performance, employee, analytics, covid-19 (key words**)**

I. Introduction

The Covid-19 pandemic has turned the work world upside down. In just a year, we've seen a massive shift to remote and hybrid work. This big change has made keeping talented employees a real challenge. The old ways of keeping staff happy and engaged, which relied a lot on face-to-face interactions, just don't cut it anymore in this new landscape.

There's a pressing need for fresh, data-smart ways to understand why people leave their jobs and how to keep them around in this new remote work reality.

Employee turnover has always been a major headache for companies. When good people leave, it hits productivity, team morale, and the company's bottom line. But figuring out why people quit has always been tricky, especially in today's diverse and fast-changing workplaces. With so many people working from home now, it's even harder to spot the warning signs that someone might be thinking of leaving

In this project, we're going to dig deep into employee data to uncover hidden patterns about why people leave their jobs. We'll use some clever data science techniques to look at the information in new ways. It's like being a detective, but instead of solving crimes, we're solving the mystery of employee attrition.

We'll use some smart computer programs to predict who might be at risk of leaving. These tools are great at spotting complex patterns that humans might miss. We'll try out different methods, like Logistic Regression and Random Forest, to see which one is best at predicting who might quit based on various factors we uncover.

To make all this data easy to understand, we'll create lots of eye-catching charts and graphs. These visuals will help bosses and HR teams quickly grasp what's going on and take action to keep their best people.

The main goal here is to really understand what makes people want to stay or leave their jobs, especially when they're working from home. By doing this, we hope to help companies keep their talented employees happy and productive, even in these new and challenging times. This project isn't just about numbers and charts – it's about helping businesses adapt to the new world of work and keep their most valuable asset: their people.

II. Previous Work

Human interest has increased in applying machine learning algorithms to predict employee performance over the past years. A series of studies explores diverse techniques and approaches to understand and improve the accuracy of predictive models.

A major study [1] also noted the difficulties in applying machine learning models to assess employee performance. This emphasizes the fact that comprehensive and interpretable models need to be developed, which are robust in handling the complexities and nuances embedded in employee performance data. Paul's work underlines the importance of generalization from different datasets and features into organizational contexts.

In a similar study [2], conducted an in-depth analysis of the performance of contract employees using five different machine learning models: Decision Tree, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, and Random Forest. It shows that the Random Forest algorithm obtained the highest performance, with superior accuracy and robustness compared to other models. Siahaan's study proves the effectiveness of ensemble methods particularlly Random Forest in handling high-dimensional and complex datasets, making it a preferred method in predictive modeling within HR analytics.

In yet another crucial research, [3] compared the prediction of employee performance by three important algorithms: Logistic Regression, Decision Tree, and Artificial Neural Network. The study found that ANN provided better classification accuracy than other models for predicting employee performance. Adeniyi's work reflects the capability of deep learning techniques in identifying complex patterns present in the data, which traditional methods often miss out on.

In another paper the authors [4] extended this research by using six different machine learning algorithms to optimize and predict employee performance. These was Logistic Regression, Decision Tree, Random Forest, Gradiant Boosting Machine (GBM), XGBost and the Support Vector Machine (SVM). Her study also showed that Random Forest consistently outperformed all the other models, confirming the earlier work of Siahaan [2]. The research emphasized feature selection and engineering to enhance the performance of the models. By selecting and transforming features carefully, Tanasescu managed to upgrade the predictive power of the models.

Together, these studies emphasize the predictive potential of machine learning with regard to employee performance but also point out the challenges and important considerations required for model building. Importantly, the selected algorithm, chosen features, and steps of data pre-processing are critical factors that largely determine the performances of the predictive models.

III. Methodology

This study employed a comprehensive approach to analyze employee attrition using HR analytics data [5]. The methodology includes data preprocessing, exploratory data analysis (EDA), and the development of a predictive model using XGBoost with hyperparameter optimization.

The initial phase involved data preparation and preprocessing using python on Jupyter notebook as the IDE. The HR Analytics from Kaggle went through several transformations. The employee ID was removed as it was not relevant for analysis. The target variable, 'Attrition', was converted to a binary numeric format (1 for 'Yes', 0 for 'No'). To handle categorical variables, one-hot encoding was applied using scikit-learn's OneHotEncoder, transforming categorical data into a format suitable for machine learning algorithms. The encoded features were then combined with the numeric features to create a comprehensive dataset for analysis.

Exploratory Data Analysis (EDA) was conducted to gain insights into the relationships between variables. A correlation matrix was done to identify the strengths of associations between different features. Particular attention was paid to the correlations with 'PerformanceRating' and 'Attrition', as these were key variables of interest. The top 10 features most strongly correlated with each of these variables were identified and examined.

We used the insights from the correlation matrix to create our visualizations. For this task we used python, specifically the libraries of seaborn and matplotlib. First we applied Trellis Plot on “JobSatisfaction” followed by “AgeGroup” and “Attrition”. Then we applied stacked bar plot on “YearsScienceLastPromotion” and “Attrition”. A Box plot on “PerformanceCategory” VS “percentSalaryHike” Attribites. Followed by a Bar plot on “Attrition” VS “OverTime”. Finally at the end a grouped bar plot on “MaritalStatus” and “StockOptionLevel” attributes. By applying visualizations on these attributes we gained some useful insights of the dataset.

To address the potential issue of class imbalance in the predictive modeling, which is common in attrition prediction tasks, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This technique creates synthetic examples of the minority class, balancing the dataset and potentially improving model performance.

Feature scaling was performed using StandardScaler to normalize the feature set, ensuring all variables were on a comparable scale. This is particularly important for algorithms sensitive to the scale of input features, such as the XGBoost algorithm used in this study.

The modeling approach used XGBoost, a powerful gradient boosting algorithm known for its effectiveness in various machine learning tasks. To optimize the model's performance, Optuna, a hyperparameter optimization framework, was used. An objective function was defined to tune key XGBoost parameters such as max_depth, learning_rate, n_estimators, min_child_weight, subsample, colsample_bytree, gamma, and scale_pos_weight. The optimization process used cross-validation with an F2 score as the evaluation metric, placing a higher emphasis on recall over precision, which is often desirable in attrition prediction cases.

After obtaining the optimal hyperparameters, the final XGBoost model was trained on the entire training set. The model's performance was evaluated on a held-out test set, which is 20% of the original data. To fine-tune the classification threshold, a Receiver Operating Characteristic (ROC) curve analysis was performed. The optimal threshold was determined by maximizing Youden's J statistic, which balances sensitivity and specificity.

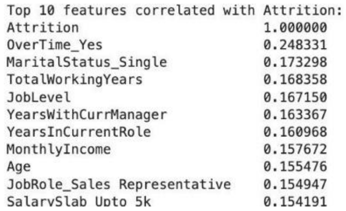
The model's performance was assessed using multiple metrics including accuracy, precision, recall, and F1-score. Additionally, a confusion matrix was generated to provide a detailed breakdown of the model's predictions, allowing for a comprehensive understanding of its strengths and potential areas for improvement.

This methodology combines robust data preprocessing, exploratory analysis, advanced modeling techniques, and thorough evaluation to create a comprehensive approach to employee attrition prediction. The use of SMOTE for handling class imbalance, XGBoost for modeling, and Optuna for hyperparameter optimization represents a comprehensive approach to this critical HR analytics task.

IV. Results

We have divided the result section in three separate sections,

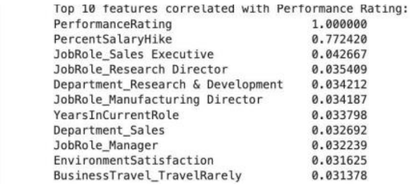
A. EDA



```
Top 10 features correlated with Attrition:
Attrition                        1.000000
OverTime_Yes                     0.248331
MaritalStatus_Single             0.173298
TotalWorkingYears                0.168358
JobLevel                         0.167150
YearsWithCurrManager             0.163367
YearsInCurrentRole               0.160968
MonthlyIncome                    0.157672
Age                              0.155476
JobRole_Sales Representative     0.154947
SalarySlab_Upto 5k               0.154191
```

Table.1 : Top 10 Features correlated with attrition.

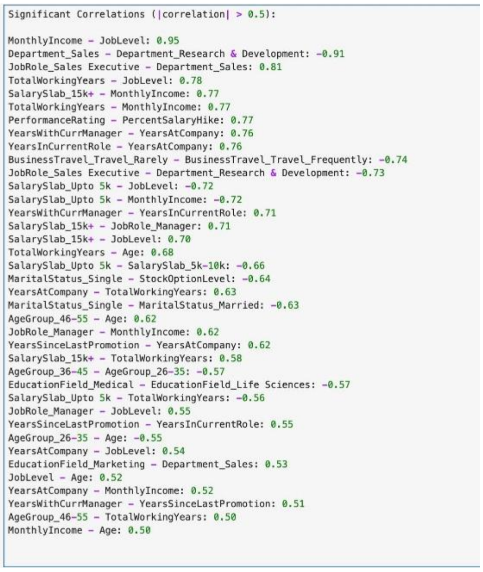
From the correlation analysis, we can see that the strongest predictor of attrition is working overtime, with a correlation of 0.248. Other notable factors positively correlated with attrition include being single (0.173), total working years (0.168), job level (0.167), and years with current manager (0.163).



```
Top 10 features correlated with Performance Rating:
PerformanceRating                        1.000000
PercentSalaryHike                        0.772420
JobRole_Sales Executive                  0.042667
JobRole_Research Director                0.035409
Department_Research & Development        0.034212
JobRole_Manufacturing Director           0.034187
YearsInCurrentRole                       0.033798
Department_Sales                         0.032692
JobRole_Manager                          0.032239
EnvironmentSatisfaction                  0.031625
BusinessTravel_TravelRarely              0.031378
```

Table.2 : Top 10 features correlated with performance rating.

There is a very strong positive correlation (0.772) between performance rating and percent salary hike. This indicates that higher-performing employees tend to receive larger salary increases.



```
Significant Correlations (|correlation| > 0.5):

MonthlyIncome - JobLevel: 0.95
Department_Sales - Department_Research & Development: -0.91
JobRole_Sales Executive - Department_Sales: 0.81
TotalWorkingYears - JobLevel: 0.78
SalarySlab_15k+ - MonthlyIncome: 0.77
TotalWorkingYears - MonthlyIncome: 0.77
PerformanceRating - PercentSalaryHike: 0.77
YearsWithCurrManager - YearsAtCompany: 0.76
YearsInCurrentRole - YearsAtCompany: 0.76
BusinessTravel_Travel_Rarely - BusinessTravel_Travel_Frequently: -0.74
JobRole_Sales Executive - Department_Research & Development: -0.73
SalarySlab_Upto 5k - JobLevel: -0.72
SalarySlab_Upto 5k - MonthlyIncome: -0.72
YearsWithCurrManager - YearsInCurrentRole: 0.71
SalarySlab_15k+ - JobRole_Manager: 0.71
SalarySlab_15k+ - JobLevel: 0.70
TotalWorkingYears - Age: 0.68
SalarySlab_Upto 5k - SalarySlab_5k-10k: -0.66
MaritalStatus_Single - StockOptionLevel: -0.64
YearsAtCompany - TotalWorkingYears: 0.63
MaritalStatus_Single - MaritalStatus_Married: -0.63
AgeGroup_46-55 - Age: 0.62
JobRole_Manager - MonthlyIncome: 0.62
YearsSinceLastPromotion - YearsAtCompany: 0.62
SalarySlab_15k+ - TotalWorkingYears: 0.58
AgeGroup_36-45 - AgeGroup_26-35: -0.57
EducationField_Medical - EducationField_Life Sciences: -0.57
SalarySlab_Upto 5k - TotalWorkingYears: -0.56
JobRole_Manager - JobLevel: 0.55
YearsSinceLastPromotion - YearsInCurrentRole: 0.55
AgeGroup_26-35 - Age: -0.55
YearsAtCompany - JobLevel: 0.54
EducationField_Marketing - Department_Sales: 0.53
JobLevel - Age: 0.52
YearsAtCompany - MonthlyIncome: 0.52
YearsWithCurrManager - YearsSinceLastPromotion: 0.51
AgeGroup_46-55 - TotalWorkingYears: 0.50
MonthlyIncome - Age: 0.50
```

Table.3 : Top significant correlations.

Interestingly, certain job roles like Sales Executive, Research Director, and Manufacturing Director show weak positive correlations with performance ratings, suggesting these positions may have slightly higher average performance scores.

Monthly income and job level are very strongly correlated (0.95), as are total working years and job level (0.78). This indicates that higher job levels are associated with higher salaries and more work experience. Additionally, there's a strong correlation (0.77) between total working years and monthly income, emphasizing on the link between experience and compensation.Years with current manager and years at the company show a strong positive correlation (0.76), as do years in current role and years at the company. This suggests that employees who stay with the company longer tend to have longer tenures with their current managers and in their current roles. Age is moderately correlated with total working years (0.68) and job level (0.52), indicating that older employees tend to have more work experience and higher positions. There's also a negative correlation (-0.64) between being single and stock option level, suggesting that married employees may have higher stock option levels. .

These correlations provide valuable insights into the relationships between various factors in the employee dataset, highlighting patterns in attrition, performance, compensation, and career progression. However, it's important to note that correlation does not imply causation, and further analysis would be needed to determine the exact nature and reasons for these relationships.

B. Visualization

Here we applied a total of 5 visualizations based on the EDA(correlation matrix).

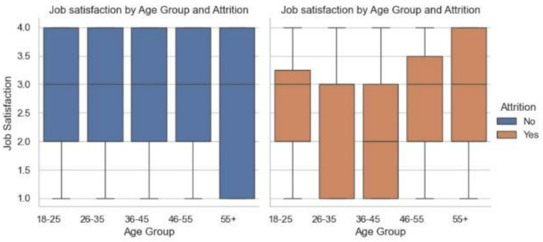


Fig.1 : Trellis Plot of Job satisfaction by Age Group and Attrition.

Among the group of employees who did not leave the company(attrition “no”); regardless of their age, most of them show job satisfaction level of between 3 and 4. While employees who left(attrition “yes”); employees between the ages of 26-35 and 36-45 show satisfaction level of between 3 or less. This suggests that, employees with higher satisfaction don’t tend to leave the company. However there might be a connection between lower job satisfaction and employee leaving the company among younger employees(ages between 26-45).

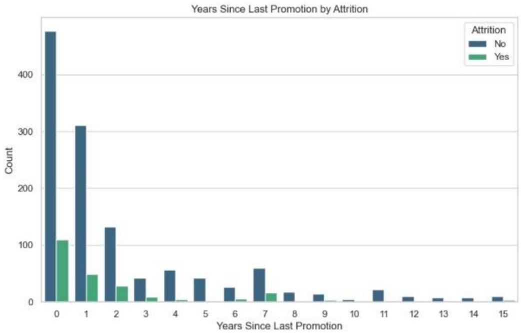


Fig.2 : Stacked Bar plot of Years Since Last promotion by Attrition.

Most of the employees who have left the company(Attrition “yes”) have 0 years since last promotion. This suggests many employees leave before their first promotion. However, if they receive promotion more than 2 years, then the rate of employees leaving the company decreases significantly. This tells us that long term employees are more sustainable.

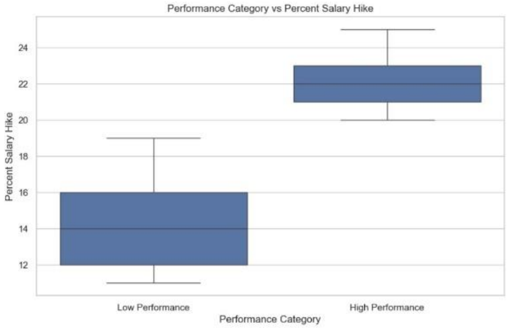


Fig.3 : Box Plot of performance Category VS percent Salary Hike.

The median percent salary hike for low performance are around 14% while for high performers it is around 22%. This indicates that employees with higher salary hikes perform better.

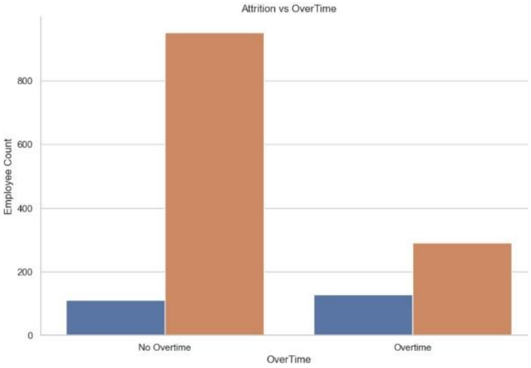


Fig.4 : Bar Plot of Attrition VS Over Time

The graph indicates that overtime is correlated with attrition. Most of the employees who did not do overtime work, stayed at the company. This could also mean that overtime can cause employees to leave the company.

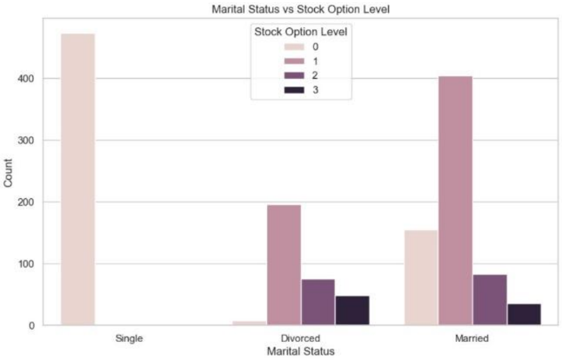
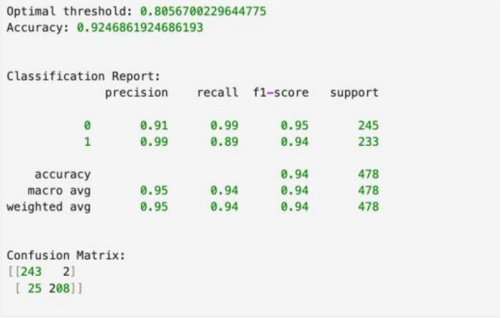


Fig.5 : Grouped Bar Plot of Marital Status VS Stock Option Level.

From the graph we can see that if the employee is single then they are not buying any stock options from the company. But among the other 2 groups(Married and Divorced), Married employees are buying stocks more compared to divorced employees. The most bought stock is level 1 stock option which is invested by mostly married employees.

C. Model



```
Optimal threshold: 0.8056700229644775
Accuracy: 0.9246861924686193


Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.99      0.95       245
           1       0.99      0.89      0.94       233

    accuracy                           0.94       478
   macro avg       0.95      0.94      0.94       478
weighted avg       0.95      0.94      0.94       478


Confusion Matrix:
[[243   2]
 [ 25 208]]
```

Table.4 : Model evaluation.

The XGBoost model achieved an overall accuracy of 92.47% on the test set, indicating that it correctly predicted employee attrition in about 9 out of 10 cases. The optimal threshold of 0.81 was determined using the ROC curve to balance precision and recall.

For non-attrition (Class 0), the model achieved 91% precision and 99% recall. This means it correctly identified 99% of employees who didn't leave, with only 1% false negatives. For attrition (Class 1), the model achieved 99% precision and 89% recall. This indicates that when the model predicted attrition, it was correct 99% of the time, but it missed about 11% of actual attrition cases.

The high F1-scores (95% for non-attrition, 94% for attrition) demonstrate a good balance between precision and recall for both classes.

Confusion Matrix Analysis:

Out of 478 test cases:

- True Negatives (243): Correctly predicted non-attrition cases.

- False Positives (2): Only 2 cases were incorrectly flagged for attrition.

- False Negatives (25): The model missed 26 actual attrition cases.

- True Positives (208): Correctly identified attrition cases.

This matrix shows that the model is particularly strong at avoiding false positives, which is important in HR applications to prevent unnecessary interventions.

The model shows signs of overfitting, with a training accuracy of 99.69% compared to a test accuracy of 92.47%. While this 7.22 percentage point difference is noticeable, the test set performance is still strong, indicating good generalization to unseen data.

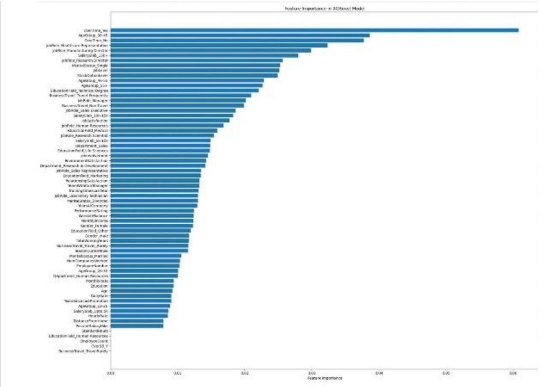


Fig.6 : XGBoost model feature important.

The feature importance analysis from the XGBoost model differs somewhat from the correlation analysis:

OverTime_Yes is the most important feature in both analyses, suggesting it's a strong predictor of attrition. Age-related features (AgeGroup_36-45) are important in the model but don't appear in the top correlations. Job roles like Healthcare Representative and Manufacturing Director are important in the model but not in the top correlations. The correlation analysis highlights factors like MaritalStatus_Single and TotalWorkingYears, which don't appear in the top model importances.

This difference suggests that while some features may not have strong linear correlations with attrition, they become important in the context of the model's complex decision trees.

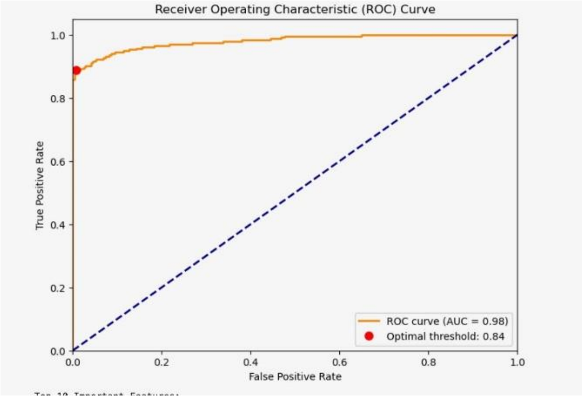


Fig.7 : ROC Curve.

The ROC curve's AUC of 0.98 is excellent, indicating that the model has a 98% chance of correctly distinguishing between attrition and non-attrition cases. This high AUC supports the model's strong performance metrics.

The model is highly accurate but slightly conservative in predicting attrition (higher false negatives than false positives). Overtime appears to be the most critical factor related to attrition risk. Age group, job role, and marital status are also important factors to consider. The model's high precision for attrition predictions (99%) means that when it flags an employee as at risk, it's very likely to be correct.

These detailed results provide a comprehensive view of the model's performance and offer valuable insights for HR strategies aimed at reducing attrition. The high accuracy and balanced performance across classes suggest that this model could be a reliable tool for identifying employees at risk of leaving, allowing for timely interventions.

V. Conclusions

We have explored the HR dataset starting with a correlation analysis to get the features that have a significant relationship with one another, and to also get the features that have a significant relationship with attrition and performance rating. Overtime came out to be the strongest predictor of attrition rating while PercentSalaryHike came out to be the strongest predictor of Performance rating. From the correlation analysis, we were able to get other significant relationships which we made visualizations for. Using trellis plot, we were able to show that the people with less job satisfaction, especially employees in the group of 26-46 years are most likely to leave the company. Using a stacked barplot, we were able to show that employees usually leave the company before getting their first promotion (within 2 years). Using a barplot, we observed that people with less overtime are most likely to stay at the company.

We also made an attrition prediction model using XGBoost. The model showed a 99% accuracy on the training set while showing a 92% accuracy on the test set. The ROC curve's AUC of 0.98 is excellent, indicating that the model has a 98% chance of correctly distinguishing between attrition and non-attrition cases. This high AUC supports the model's strong performance metrics.

The model is highly accurate but slightly cautious in predicting attrition (higher false negatives than false positives). The confusion matrix shows that the model is particularly strong at avoiding false positives, which is important in HR applications to prevent unnecessary interventions.

VI. References

[1] T. Paul, "Strategic Employee Performance Analysis in the USA: Leveraging Intelligent Machine Learning Algorithms," 08 05 2024. [Online]. Available: https://unbss.com/index.php/unbss/article/view/33.

[2] M. Siahaan, "An Analysis of Contract Employee Performance Assessment Using Machine Learning," 16 07 2021. [Online]. Available: https://ojs.uma.ac.id/index.php/jite/article/view/5357.

[3] J. a. A. A. E. a. Y. J. O. a. E. G. O. a. A. K. D. a. O. P. C. a. A. S. A. Adeniyi, "Comparative Analysis of Machine Learning Techniques for the Prediction of Employee Performance," 15 01 2024. [Online]. Available: https://eprints.lmu.edu.ng/4419/.

[4] A. V. A. R. B. a. O. V. Laura Gabriela Tanasescu, "Data Analytics for Optimizing and Predicting Employee Performance," 12 04 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/8/3254.

[5] S. HAROON, "HR Analytics Dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/saadharoon27/hr-analytics-dataset?resource=download.