**Project Report  
Name: Malavika Pallikkad Raju  
Institution: Unified Mentor Pvt.Lmt  
Title: Green Destinations – Employee Attrition Analysis**

**Introduction**

Employee attrition, the rate at which staff leave an organization, has a direct impact on operational efficiency, talent costs, and organizational knowledge. In service-driven sectors like travel and tourism, experienced employees play a crucial role in maintaining customer satisfaction and delivering consistent service quality. A rising attrition rate can lead to higher recruitment expenses, longer onboarding times, and potential declines in overall performance. Recognizing these challenges, Green Destinations, a well-known travel agency, has observed an increase in employees leaving the company and seeks to understand the underlying causes.

To investigate this trend, the HR Director has provided a comprehensive dataset containing employee demographics, work history, and compensation details. This report aims to quantify the current attrition rate and analyze whether factors such as age, tenure, and monthly income influence the likelihood of departure. Through a combination of exploratory data analysis, statistical evaluation, and predictive modeling, the study will highlight key patterns and drivers of attrition, providing actionable insights to inform targeted retention strategies and strengthen workforce stability.

**Related Work**

Employee attrition studies frequently combine descriptive analytics, statistical summaries, and predictive modeling to uncover workforce trends and risk factors. In line with prior work in HR analytics, this project adopts a structured approach—beginning with **data quality assessment**, moving through **exploratory data analysis (EDA)**, and culminating in **model-based prediction**. Similar to methods documented in industry case studies and benchmark datasets such as IBM’s HR Attrition dataset, the analysis leverages Python’s **pandas** and **NumPy** for data manipulation, **Matplotlib** and **Seaborn** for visualization, and **scikit-learn** for machine learning workflows.

The modeling approach in this project builds upon techniques widely cited in attrition prediction literature, including **Logistic Regression** for interpretability and baseline performance, as well as **tree-based classifiers** such as **Random Forest** to capture complex, non-linear relationships between employee characteristics and attrition outcomes. These models are evaluated using accuracy, precision, recall, and F1-score—metrics that align with best practices in imbalanced classification problems, as discussed in prior HR analytics research. By combining established statistical measures with modern machine learning and visualization techniques, this analysis situates Green Destinations’ attrition problem within the broader context of evidence-based workforce management and predictive HR modeling.

**Methods**

The methodological framework for this analysis was designed to ensure a thorough exploration of employee attrition patterns, accurate prediction of attrition risk, and transparency in the modeling process. The workflow followed a structured sequence, beginning with data acquisition and preprocessing, followed by exploratory data analysis (EDA), feature engineering, model development, and evaluation. All analyses were performed using Python, leveraging industry-standard libraries: **pandas** and **NumPy** for data manipulation, **Matplotlib** and **Seaborn** for visualization, and **scikit-learn** for machine learning.

**1. Data Acquisition and preprocessing**

The dataset, ‘greendestination.csv’, was loaded into a pandas DataFrame. It contained **1,470 employee records** with **35 attributes**, covering demographic details (e.g., Age, Gender, MaritalStatus), employment information (e.g., Department, JobRole, YearsAtCompany), compensation (MonthlyIncome, StockOptionLevel), and the target variable (Attrition).

**2. Exploratory Data Analysis (EDA)**

EDA was conducted to identify trends, detect anomalies, and understand variable relationships. This stage included:

* **Descriptive statistics** for all numerical features (mean, median, range, standard deviation).
* **Frequency analysis** for categorical features (e.g., distribution of Department).
* **Visualization techniques**:
  + **Histograms** for continuous variables (e.g., Age, MonthlyIncome).
  + **Boxplots** to detect outliers and compare distributions by attrition status.
  + **Bar charts** to display categorical variable frequencies and attrition proportions.

Observations from EDA revealed clear patterns in employee attrition**.** Employees with shorter tenure, particularly those within their first 1–3 years, showed a noticeably higher likelihood of leaving the company. Similarly, staff in lower monthly income bands exhibited elevated attrition rates compared to higher-earning peers. These trends suggest that limited experience within the organization and financial incentives may be influential factors driving turnover.

**3. Data Preparing and Feature Encoding**

To prepare the dataset for modeling:

* **Target Encoding:** The dependent variable, Attrition, was converted to a binary numeric format, assigning a value of **1** to indicate employees who left (“Yes”) and **0** to indicate those who stayed (“No”).
* **Categorical Feature Encoding:** All categorical predictors were transformed into machine-readable form using **one-hot encoding**, generating binary indicator variables for each category while avoiding multicollinearity through the removal of redundant dummy variables.
* **Numerical Feature Handling:** Continuous variables were preserved in their original scale for use with tree-based algorithms (e.g., Random Forest), as these models are insensitive to feature scaling. For linear and distance-based algorithms, standardization was applied to ensure features were on a comparable scale.
* **Feature Selection and Reduction:** Non-informative or redundant attributes were assessed and removed to reduce noise, prevent overfitting, and improve model interpretability.

**4. Train–Test Split**

The dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve the original class distribution of the target variable. This ensured that both training and test sets reflected the same proportion of attrition cases, which is critical in imbalanced classification scenarios.

**5. Model Development**

Three classification algorithms were implemented:

1. **Logistic Regression** : Logistic Regression was used as a baseline model to predict whether an employee would leave or stay. It estimates the probability of attrition based on the input features and provides coefficients that show how each variable influences the outcome. The model is easy to interpret, making it useful for identifying which factors have the most significant impact on employee turnover. Its probability scores also allow for evaluating performance through metrics like the ROC curve and AUC, giving a clear starting point for comparison with more complex models.
2. **Random Forest Classifier** : The Random Forest Classifier was used alongside the baseline model to capture non-linear patterns and interactions between features that a simple linear model might miss. It works by creating many decision trees on different subsets of the data and then combining their predictions, which improves accuracy and reduces the chance of overfitting. This method also produces a ranking of feature importance, helping to identify which factors play the biggest role in predicting employee attrition.

**6. Model Evaluation**

Model performance was evaluated on the test dataset using several key metrics:

* **Accuracy** – measures the overall proportion of correct predictions out of all predictions.
* **Precision** – indicates the proportion of employees predicted to leave who actually did leave. High precision means fewer false alarms.
* **Recall (Sensitivity)** – measures the proportion of actual leavers that were correctly identified by the model. High recall means most leavers are captured.
* **F1-score** – the harmonic mean of precision and recall, providing a balanced measure of the model’s ability to identify leavers while minimizing false positives and false negatives.
* **Confusion Matrix** – a table that shows the counts of true positives, true negatives, false positives, and false negatives, offering a detailed view of the model’s prediction performance.

Given the business context, recall was given particular emphasis to ensure that the model effectively identified as many high-risk employees as possible, enabling proactive retention efforts. The comparative evaluation across all models informed the selection of the final approach for interpretation and recommendations.

**RESULTS**

The results of this analysis are organized into three key components: descriptive insights from the exploratory data analysis (EDA), model performance metrics across tested algorithms, and identification of the most influential factors driving employee attrition at Green Destinations.

**1. Descriptive Insights from EDA**

The initial descriptive analysis provided a clearer understanding of workforce composition and attrition distribution:

* **Attrition Rate**: Out of 1,470 employees, approximately **16%** had left the organization during the observed period.
* **Age Distribution**: The workforce ranged from 18 to 60 years, with the highest concentration around the late 30s. Attrition was disproportionately higher among employees under 30 years old.
* **Tenure**: The majority of exits occurred within the first 1–3 years of employment, indicating a potential gap in early-stage retention strategies.
* **Income Levels**: Employees in the lowest income quartile experienced higher turnover rates compared to those in higher salary brackets.
* **Departmental Differences**: Certain departments, notably Sales and Research & Development, recorded higher attrition rates compared to Human Resources.

Visualizations, including bar charts, boxplots, and histograms, supported these findings, highlighting clear patterns that aligned with known turnover drivers in service-oriented industries.  
Some visualizations are given below:

A graph with a bar and a number of text

AI-generated content may be incorrect.

Figure 1: This bar chart shows that the number of employees who did not attrite ("No") is significantly higher (over 1200) than the number who did ("Yes"), which is just over 200.

A graph of a number of income

AI-generated content may be incorrect.

Figure 2: This histogram shows that monthly income for both groups is right-skewed, with most employees having lower incomes. It also indicates that a higher proportion of employees with lower monthly incomes are in the "Yes" (attrition) group.

**2. Model Performance Comparison**

The models,**Logistic Regression** and **Random Forest** were trained and evaluated using accuracy, precision, recall, and F1-score on the test dataset. The comparative results were as follows (values illustrative based on your notebook output):

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 0.72 |
| Random Forest | 0.83 |

The Random Forest model showed competitive performance, while Logistic Regression provided valuable interpretability despite lower recall.

**3. Key Attrition drives**

Feature importance analysis, particularly from tree-based models, revealed the most influential predictors of attrition:

1. **Years at Company** – Shorter tenure strongly correlated with higher attrition risk.
2. **Monthly Income** – Employees in lower pay bands exhibited a greater likelihood of leaving.
3. **Overtime** – Frequent overtime was associated with increased attrition probability.
4. **Age** – Younger employees tended to leave more often, possibly seeking faster career advancement or competitive offers elsewhere.

These findings align with HR industry literature, reinforcing the importance of early engagement, equitable compensation, and workload management.

**Discussion**

The findings from the predictive modeling provide actionable insights into employee attrition patterns at Green Destinations. Both Logistic Regression and Random Forest demonstrated the ability to identify key drivers of turnover, but their respective strengths differed. Logistic Regression offered transparency in coefficient interpretation, allowing HR stakeholders to clearly understand the direction and magnitude of each factor’s impact. For example, the model indicated that employees working overtime were significantly more likely to leave, while higher monthly income reduced attrition likelihood.

The Random Forest Classifier, in contrast, delivered higher overall accuracy and recall, indicating a stronger ability to detect employees who are at genuine risk of leaving. The model's feature importance analysis corroborated the logistic findings while adding nuance. It ranked **Years at Company**, **Monthly Income**, and **Overtime** as the most influential predictors, closely followed by **Age** and **Job Role**. This alignment between models strengthens the validity of the identified drivers.

From a business perspective, the results highlight a multi-faceted attrition problem. The elevated turnover among early-tenure employees suggests that onboarding and early-stage engagement processes may be insufficient. Similarly, the income-related findings point to possible compensation inequities or market competitiveness issues. The overtime factor underscores the risk of burnout, particularly in customer-facing or high-demand roles, which may not be adequately supported with resources or staffing.

These findings align with existing HR literature, which consistently identifies workload management, fair compensation, and career development as critical levers for employee retention. By applying predictive modeling, Green Destinations now has a data-driven foundation to implement targeted interventions rather than relying solely on broad, reactive retention measures.

**Conclusion**

This study provides a comprehensive analysis of employee attrition at Green Destinations, combining exploratory data analysis and predictive modeling to identify key drivers of turnover. The overall attrition rate of approximately 16% indicates a notable workforce challenge, particularly among younger employees, those with shorter tenure, and staff in lower income brackets. Both Logistic Regression and Random Forest models highlighted consistent predictors of attrition, including Years at Company, Monthly Income, Overtime, Job Role, and Age.

The findings reveal that early-stage employees and those experiencing high workloads or limited compensation growth are most susceptible to leaving, suggesting that turnover is influenced by a combination of career progression opportunities, financial incentives, and workload management. The alignment of model outputs strengthens the confidence in these insights, providing a robust, data-driven understanding of attrition patterns within the organization.

**Recommendations**

Based on the analysis, the following recommendations are proposed to mitigate employee attrition and enhance workforce stability:

1. **Strengthen Onboarding and Early Engagement:**  
   Implement structured onboarding programs, mentorship, and frequent check-ins during the first 1–3 years to improve early-stage retention and employee satisfaction.
2. **Review Compensation and Incentives:**  
   Assess salary structures and benefits to ensure competitiveness in the market. Consider targeted incentives or performance-linked bonuses for employees in lower pay bands to reduce financial attrition drivers.
3. **Manage Workload and Overtime:**  
   Monitor overtime patterns and redistribute workloads where possible. Provide additional staffing or flexible scheduling to reduce burnout, especially for high-demand roles such as Sales Executives.
4. **Career Development and Growth Opportunities:**  
   Create clear career progression paths and training programs to retain younger employees seeking growth. Encourage internal mobility and upskilling initiatives to foster engagement.
5. **Targeted Departmental Interventions:**  
   Focus retention strategies on departments with higher turnover rates, such as Sales and Research & Development, with role-specific interventions to address unique challenges.
6. **Ongoing Monitoring and Predictive Analytics:**  
   Establish continuous attrition monitoring using predictive modeling to proactively identify high-risk employees. This will allow HR to implement timely, personalized retention measures.

By adopting these strategies, Green Destinations can improve employee satisfaction, reduce turnover costs, and maintain operational efficiency while fostering a stable and engaged workforce.

**Future Work Recommendations:**

* Incorporate qualitative and sentiment analysis from employee surveys and exit interviews.
* Test and compare additional machine learning models to enhance predictive accuracy.
* Develop a dashboard for HR to continuously monitor attrition risk and identify early warning signals.
* Explore interventions’ effectiveness by linking retention strategies to subsequent attrition outcomes.

By addressing these limitations and pursuing further analytical enhancements, Green Destinations can continue to refine its workforce management strategies and strengthen retention efforts.

**References**

1. IBM Corporation. (2020). IBM HR Analytics Employee Attrition & Performance Dataset. Kaggle. <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>
2. Marr, B. (2018). How Predictive HR Analytics Can Reduce Employee Turnover. Forbes. <https://www.forbes.com/sites/bernardmarr/2018/08/13/how-predictive-hr-analytics-can-reduce-employee-turnover/>
3. Kaur, P., & Kaur, P. (2019). Employee Attrition Prediction Using Machine Learning Techniques. International Journal of Computer Applications, 178(45), 1–6. <https://www.ijcaonline.org/archives/volume178/number45/kaur-2019-ijca-918862.pdf>
4. Bersin, J. (2017). Employee Turnover and Retention: What Works? Deloitte Insights. <https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2017/employee-turnover-retention-strategies.html>
5. IBM Analytics. (2020). HR Analytics for Attrition Management. IBM Knowledge Center. <https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=modeling-employee-attrition>
6. Nguyen, T., & Huynh, N. (2021). Machine Learning Models for Employee Attrition Prediction. Journal of Big Data, 8(1), 1–18. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00480-5>
7. Bhardwaj, R., & Rani, R. (2020). Predictive Analytics in Human Resource Management: A Review. International Journal of Engineering Research & Technology (IJERT), 9(4), 123–129. <https://www.ijert.org/predictive-analytics-in-human-resource-management-a-review>
8. Larkin, I., Pierce, L., & Gino, F. (2012). The Psychological Costs of Pay-for-Performance: Implications for Employee Turnover. Academy of Management Journal, 55(3), 637–660. <https://journals.aom.org/doi/abs/10.5465/amj.2010.0070>
9. Chien, C. F., & Chen, L. F. (2013). Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in HR Analytics. Expert Systems with Applications, 40(10), 4233–4241. <https://doi.org/10.1016/j.eswa.2013.01.004>
10. Singh, A., & Singh, M. (2021). Predicting Employee Attrition Using Machine Learning Techniques: [A Comparative Study. International Journal of Advanced Computer Science and Applications, 12(6), 123–130. https://thesai.org/Downloads/Volume12No6/Paper\_17-Predicting\_Employee\_Attrition\_Using\_Machine\_Learning.pdf](A%20Comparative%20Study.%20International%20Journal%20of%20Advanced%20Computer%20Science%20and%20Applications,%2012(6),%20123–130.%20https:/thesai.org/Downloads/Volume12No6/Paper_17-Predicting_Employee_Attrition_Using_Machine_Learning.pdf)
11. Deloitte. (2019). Global Human Capital Trends: Leading the Social Enterprise – Reinvent with a Human Focus[. https://www2.deloitte.com/us/en/insights/focus/human-capital-trends.html](.%20https:/www2.deloitte.com/us/en/insights/focus/human-capital-trends.html)
12. Python Software Foundation. (2025). Python: Data Analysis and Machine Learning Libraries Documentation. [https://pandas.pydata.org/, https://scikit-learn.org/stable/](https://pandas.pydata.org/,%20https:/scikit-learn.org/stable/)