Bootstrap Analysis - Bag of Little Bootstrap

GitHub

By

Rohit Kulkarni (rk1169),

Naga Satya Dheeraj Anumala (na945),

Krit Gupta (ksg124),

Sai Adarsh Kasula (sk2837)

Course: 16:958:588:01 – Data Mining

Department of Statistics & Data Science

Rutgers University

New Brunswick, NJ

Supervised by

Prof. Koulik Khamaru and Prof. Gemma Moran

Abstract

In this project, we implemented and compared the Bag of Little Bootstraps (BLB) technique with the standard bootstrap method to evaluate the efficiencies of various machine learning models. The analysis was conducted across three regression models and three classification models, utilizing distinct datasets for each category. Both bootstrapping methods were applied over 10,000 iterations to each model. The primary objective was to assess the computational efficiency and accuracy of BLB in comparison to traditional bootstrapping. Results from this comparative analysis highlight the performance variations and potential advantages of using BLB, particularly in scenarios involving large datasets. The findings contribute to a deeper understanding of bootstrap methodologies in the context of enhancing machine learning model reliability.

1. Introduction

Traditionally, Bootstrap regression techniques have been employed to evaluate the variability and confidence in statistical estimates derived from predictive models. The standard bootstrap method, which involves resampling the original data with replacement to create multiple simulated samples, is widely recognized for its robustness. However, this method can be computationally intensive, particularly as the volume of data grows, often making it impractical for large-scale applications. This computational burden can delay decision-making processes and limit the usability of Bootstrap techniques in real-time data analysis scenarios.

To address these challenges, this project focuses on the implementation and comparison of the traditional Bootstrap method and the Bag of Little Bootstraps (BLB). The BLB method is a refined approach specifically designed to handle the demands of large datasets. It modifies the classic Bootstrap by introducing a strategy of subsampling, resampling, scaling, and aggregating estimates, which significantly reduces computational load while maintaining estimation accuracy.

The primary aim of this project is to investigate whether the BLB can provide a more computationally efficient yet equally reliable alternative to the Standard Bootstrap method in the context of large datasets. By applying both techniques to various regression and classification models, this study will assess their performance in terms of computational efficiency, accuracy of uncertainty measures, and scalability. The outcomes of this comparative analysis highlight the practical benefits of BLB, especially its capability to provide quick and robust statistical inferences, which is crucial for timely and informed decision-making in data-intensive fields.

This project addresses the pressing need for efficient and reliable methods to assess the stability of predictive models in big data settings. Through a detailed exploration and evaluation of the BLB method compared to the traditional Bootstrap approach, this research aims to contribute valuable insights and practical solutions that enhance the application of statistical estimation techniques.

2. Materials and Methods

In addressing the problem of evaluating the stability and reliability of predictive models in large datasets, this study employed rigorous data analysis methods and advanced statistical techniques. By integrating robust bootstrap methods into our model evaluation framework, we aimed to investigate the effectiveness of traditional Bootstrap and the Bag of Little Bootstraps methods in providing reliable uncertainty measures.

Below, is outline the comprehensive approach taken from initial data handling to the processes of model training and evaluation, ensuring that each phase of the methodology contributes effectively towards solving the outlined problem.

2.1 Data Preprocessing and Setup

This project utilizes two distinct datasets, one for Regression Analysis and one for Classification. For regression analysis, this project uses the Zomato dataset – an online ordering food portal, which contains information about various restaurants, including their location, cuisine types, ratings and the typical cost for a meal for two.

For classification analysis, data regarding employee performance and promotion status were analyzed from the HR Analytics dataset with the goal to predict whether employees are promoted or not, based on features such as performance ratings, years of experience, and training scores.

2.2 Regression Analysis – Zomato

2.2.1 Data Overview

The dataset used for the regression analysis provides comprehensive details about restaurants listed on Zomato. This dataset was chosen for its rich attributes, allowing us to explore various factors influencing the dining choices and preferences among users. The key details captured in the dataset include:

1. **Restaurant Info**: Basic information such as the name, address, location, and contact details of the restaurants.
2. **Services**: Attributes indicating whether the restaurant offers online ordering or table booking services.
3. **Customer Ratings and Reviews**: Quantitative and qualitative measures of customer satisfaction and popularity.
4. **Cuisines**: Various types of cuisines offered by the restaurants, reflecting the diversity in food choices.
5. **Restaurant Type**: Classifications like ‘Cafes’, ‘Delivery’, ‘Dine-out’, and more, providing insights into the service modes.
6. **Costs**: The average cost for two people, serving as the target variable, this reflects the affordability and price segment of each restaurant.

2.2.2 Exploratory Data Analysis

Exploratory Data Analysis was conducted to uncover trends, patterns, and relationships within the Zomato dataset. Through a series of visualizations, insights were extracted that served as a prelude to predictive modeling. Here are the synthesized observations:

1. Missing Data Analysis

The initial step in the analysis involved assessing the completeness of the dataset by identifying missing values across various attributes. The results indicated several key points for consideration:

a. **Significant Missing Data**: The ***'rate'*** column, crucial for understanding customer satisfaction, had 7,775 missing entries. The ***'dish_liked'*** column had the highest number of missing values at 28,078, suggesting a significant gap in specific data entries which could affect the depth of analysis related to customer preferences.

b. **Moderate Missing Values**: Columns such as ***'phone'*** and ***'rest_type'*** had moderate amounts of missing data, which could potentially impact the accuracy of any business insights derived from these attributes. These columns were **dropped** from the dataset.

c. **Financial Data** Integrity: The ***'approx_cost(for two people)'*** column, crucial for economic analysis, had 346 missing values, which necessitated immediate cleaning to ensure the financial analysis's reliability. These missing data points were **dropped**.

2. **Online Service Trends:** From Fig. 2.1, it can be observed that a predominant portion of restaurants, facilitate **online orders**, pointing to the digital inclination of the food industry. However, a more traditional approach is seen in table reservations, with many establishments not offering this service through the platform.

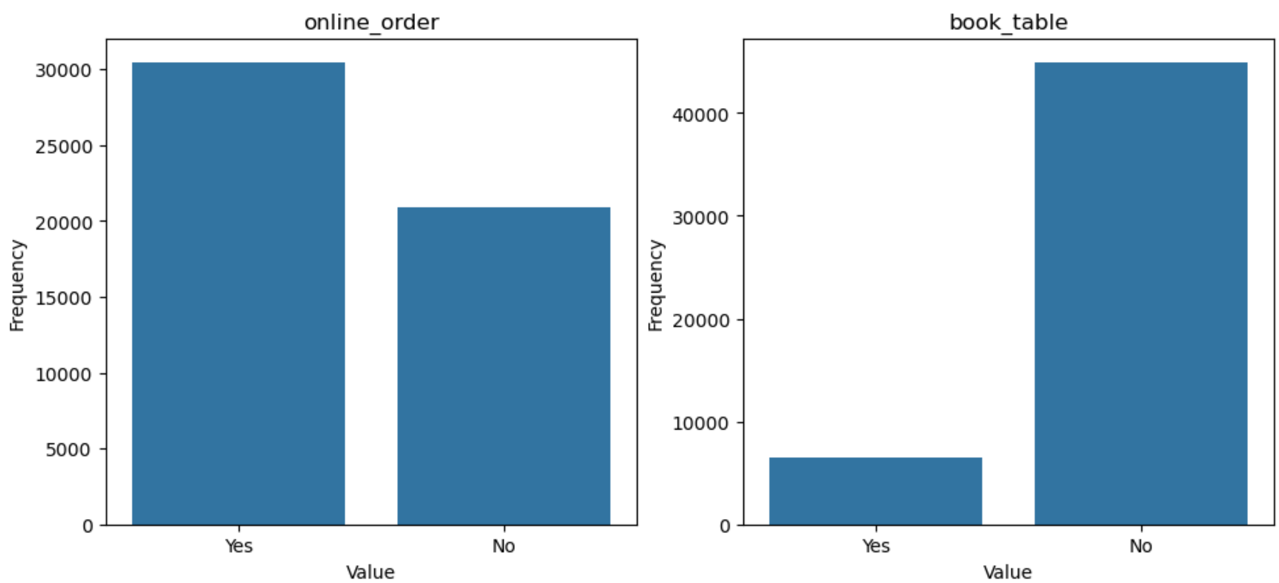

Figure 2.1. Distribution of Online Order and Reservation Status

3. **Service Type Preferences:** Fig. 2.2 provides illustrates the distribution of restaurant types. It can be observed that **'Delivery'** dominates the market, with **'Dine-out'** options also being significantly represented, indicating consumer preference for an on-the-go dining experiences.'

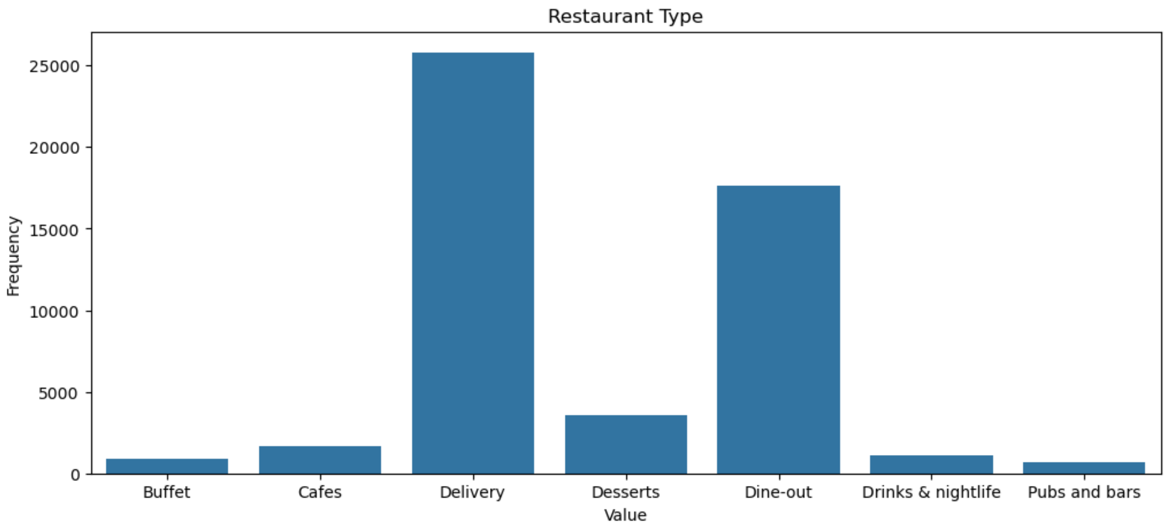

Figure 2.2. Distribution of Restaurant Types

4. **Geographic Restaurant Distribution:** From Fig. 2.3, it can be observed that the geographical distribution of restaurant is somewhat **uniform**, with a concentration of restaurants in certain neighborhoods - suggesting regional hotspots.

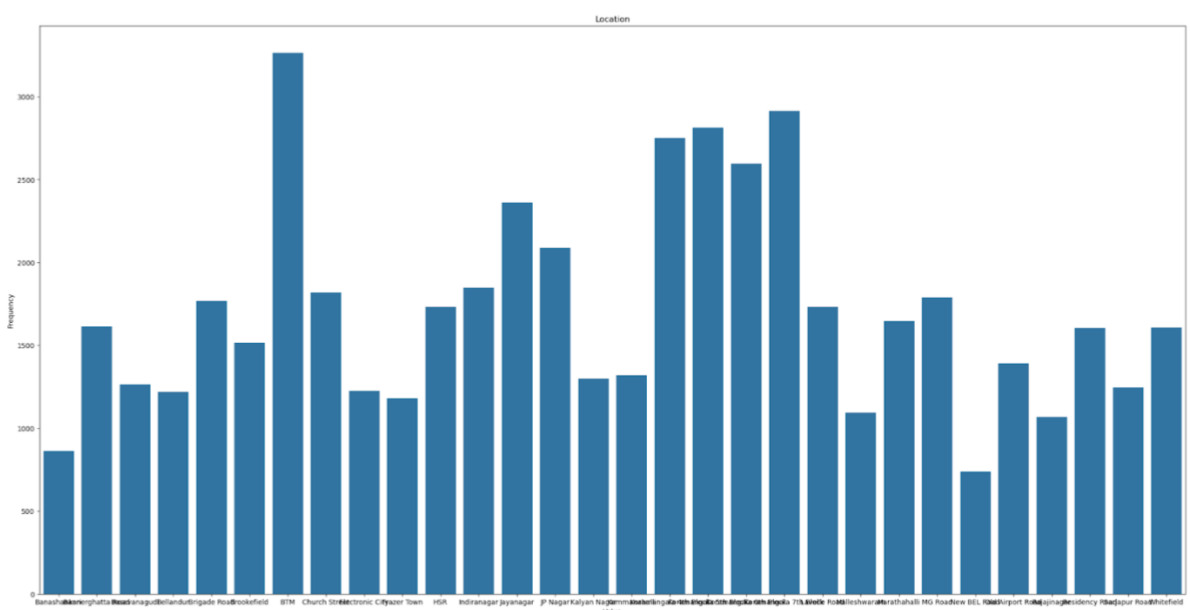

Figure 2.3. Distribution of Restaurant Geographies

5. **Customer Satisfaction Gauge:** From Fig. 2.4, The histogram of ***rating*** exhibits a left-skewed distribution, mostly leaning towards the higher end of the scale, suggesting that customers tend to rate restaurants favorably on Zomato.

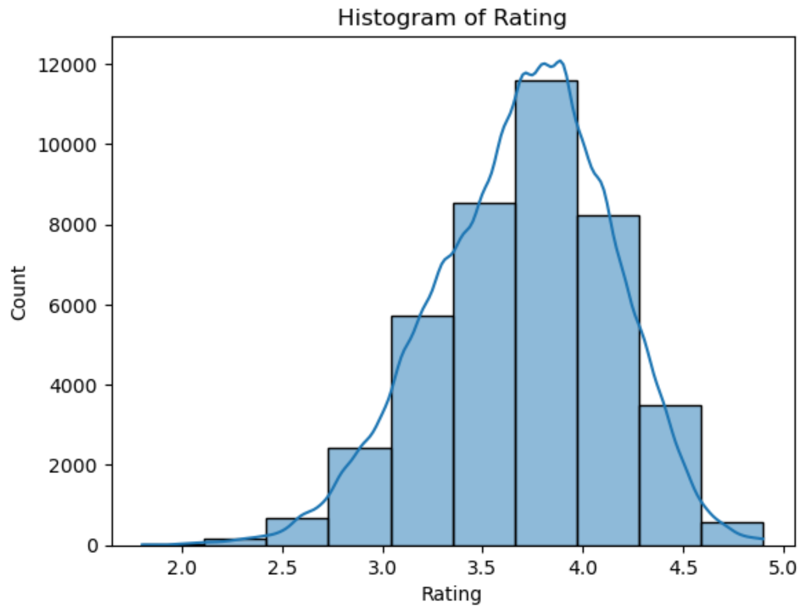

Figure 2.4. Distribution of Rating (with KDE)

6. **Average Customer Cost Analysis**: From Fig. 2.5, a notable skew in the average cost was observed in the cost distribution among restaurants.

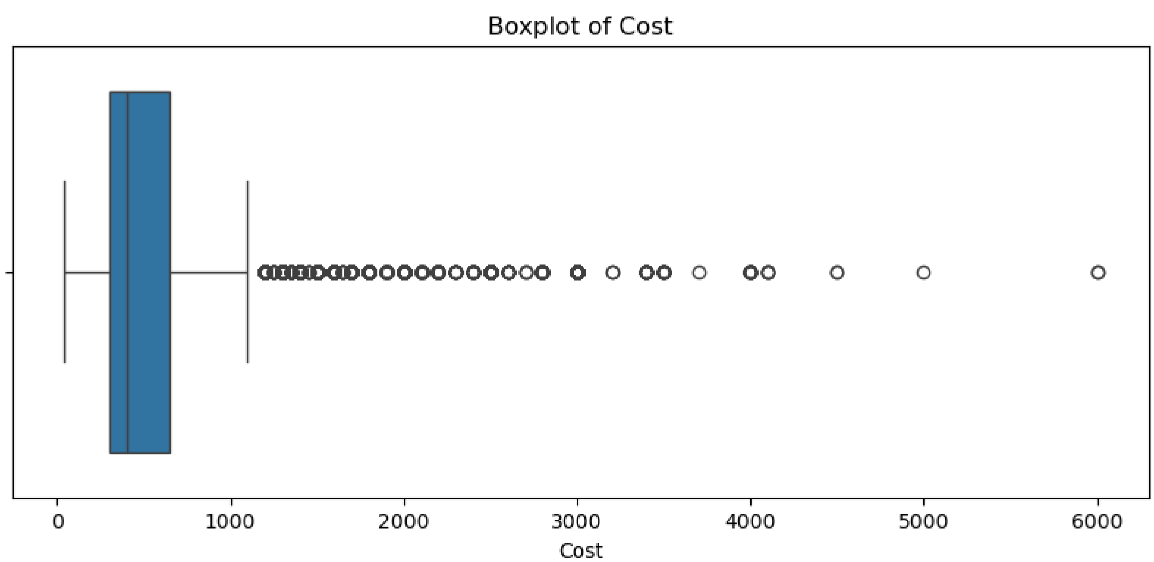

Figure 2.5. Distribution of Average Cost for 2 people

To address this, a log transformation was applied, which helped effectively **reduced skewness**, with only a few remaining outliers, as seen in Fig. 2.6.

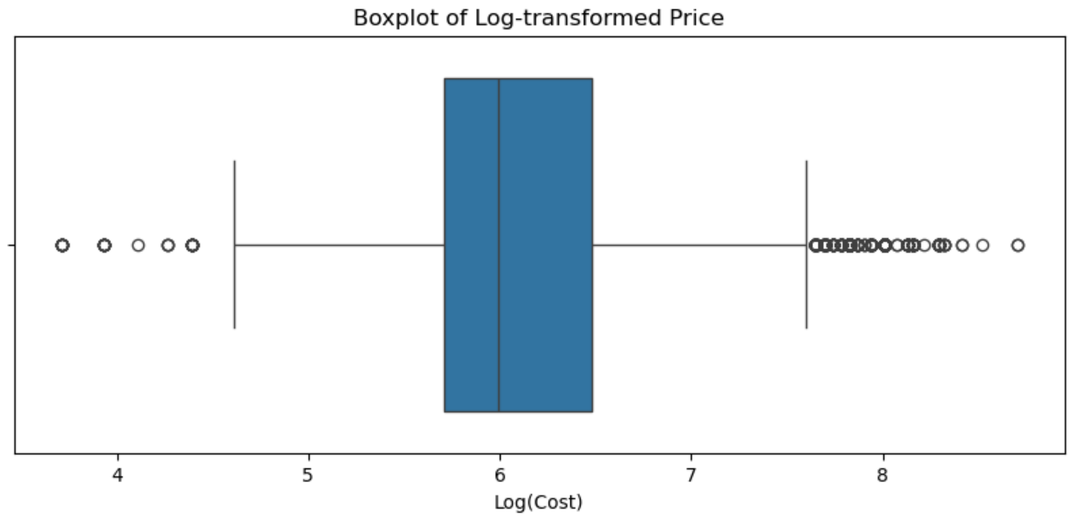

Figure 2.6. Distribution of Log-transformed Average Cost for 2 people

To address this, Winsorization was employed to limit extreme values from both ends of the distribution, effectively reducing the influence of outliers. This involves replacing the most extreme data points with the nearest values that are less extreme, which helps in stabilizing the dataset. The boxplot of the winsorized log costs confirms a more uniform distribution, indicating that the impact of outliers on the analysis is minimized, as seen in Figure 2.7.

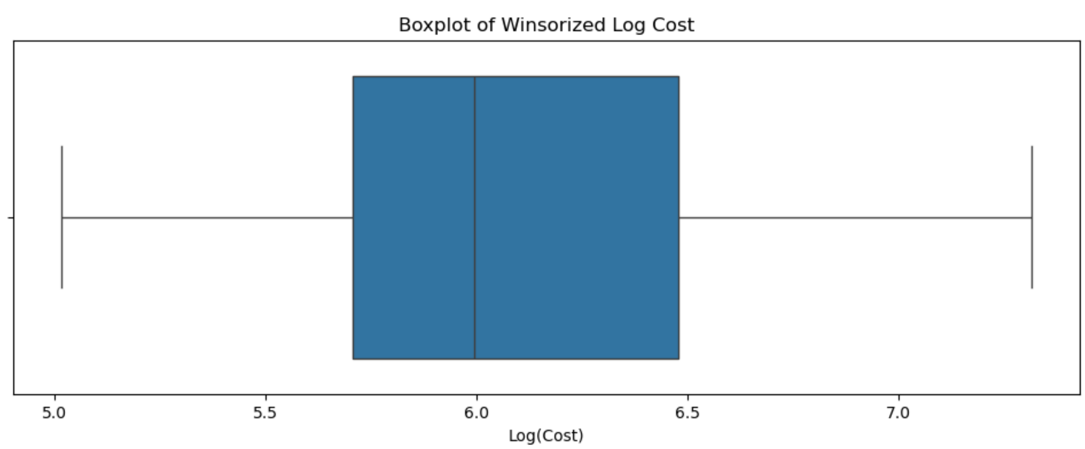

Figure 2.7. Boxplot of Winsorized Log Average Cost for 2 people

2.2.3 Feature Engineering

2.2.3.1 Data Cleaning and Transformation:

To enhance the dataset's quality, initially:

1. **Missing Cost Data Handling**: Entries where ***'approx_cost(for two people)'*** was missing were removed, ensuring that the analysis would be based on complete and reliable data.
2. **Cost Data Normalization**: The cost data was transformed from a string format, which included commas, to a numerical format. This transformation is crucial for subsequent financial analysis and modeling.
3. **Removing Unnecessary Columns**: Columns which either had no relevance to the performance of the model or had missing values >50% in the data entries were dropped.
4. **Encoding Categorical Variables**: The dataset features categorical variables that are transformed into a format suitable for machine learning analysis using Label encoding.

2.2.3.3 Vectorization of the cuisine data

The dataset presented a wide array of cuisines, which are qualitative and categorical in nature. To incorporate this richness of data into predictive models, vectorization was used to transform the 'cuisines' feature. By employing a Word2Vec model, the list of cuisines was converted into numerical vectors, capturing the semantic relationships between different cuisines, which encapsulates the variety of cuisines into a format amenable to machine learning algorithms.

2.2.3.4 Imputation of Missing Ratings

Given the substantial number of missing entries in the ***'rate'*** column, an advanced imputation technique was employed to estimate these values, enhancing the dataset's completeness and usability. The **Multiple Imputation by Chained Equations** (**MICE**) technique was utilized, leveraging an iterative version of the imputer with a Linear Regression estimator to predict missing ratings based on other available data. This method is particularly effective for handling missing data by using the relationship between features to estimate missing values accurately. Engineered features were integrated with other numerical features that encapsulate the type of cuisine in numeric form. The iterative imputer utilized these features to predict and fill in missing ratings, ensuring that each entry in the dataset now had a complete set of data for accurate analysis. By handling missing data, the dataset is now more robust, complete, and ready for regression analysis.

Post feature engineering, a correlation heatmap was generated to understand the strength and direction of relationships between the variables in a dataset. From Fig. 2.8, it can be observed that:

1. There is a **moderate positive correlation** between ***votes*** and ***rating***, suggesting that customer engagement through voting is an indicator of satisfaction.
2. A **moderate positive correlation** exists between the ***number of votes*** and the ***cost***, implying that restaurants with more votes tend to be pricier.
3. A similar **moderate positive correlation** was observed between the ***rating*** and ***cost***, suggesting that higher-rated restaurants might also have higher costs.

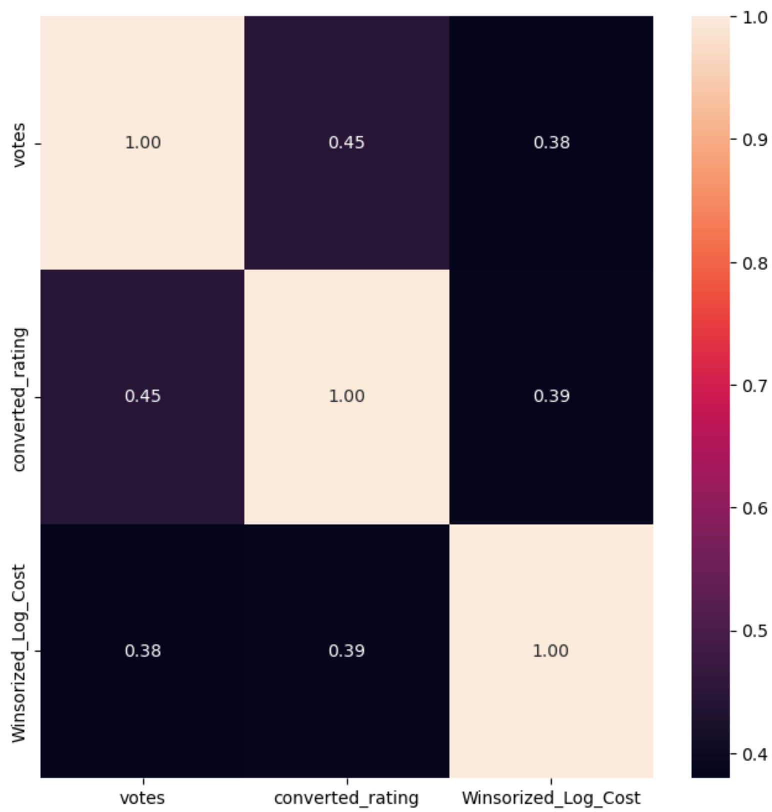

Figure 2.8. Heatmap of Numerical Features

2.3 Classification Analysis – HR Analytics

2.3.1 Data Overview:

The classification dataset consists of information about various employees who are candidates for promotion within an organization. This dataset captures a broad spectrum of performance and personal attributes, making it ideal for predicting potential promotions based on empirical data.

Key attributes include:

1. **Personal Info**: Details such as Employee ID, Gender, Age, and educational background.
2. **Employment Details**: Information about the employee's department, region, recruitment channel, and length of service.
3. **Performance Metrics:** Various indicators of an employee's performance, including training attended, performance rating, whether key performance indicators (KPIs) were met more than 80% of the time, any awards won, and the average training score.
4. **Promotion Status**: This binary target variable indicates whether an employee was promoted or not, serving as a critical measure for assessing the effectiveness of the predictive models.

2.3.2 Exploratory Data Analysis:

1. Missing Data Analysis:

The ***'education'*** attribute, which is critical for understanding an employee's educational background, has **2,409** missing entries. Similarly, the ***'previous_year_rating'*** attribute, crucial for assessing past performance, has 4,124 missing entries. These gaps are significant as they directly relate to factors that could influence promotion decisions.

2. Distribution of the Target Variable:

The pie chart from Fig 2.9, the distribution of ***'promotion status'*** is observed, where 8.5% of the employees depicted were promoted, while a significant majority, 91.5%, were not promoted. This significant imbalance suggests that promotions are quite selective or potentially indicative of a high-performance threshold within the company's promotion criteria.

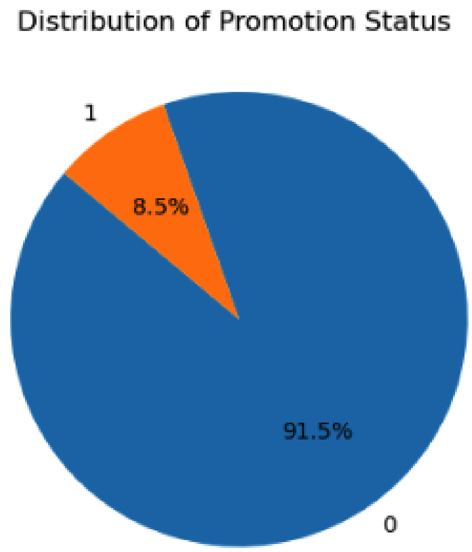

Figure 2.9. Distribution of Promotion Status

3. Distribution of Numerical Features:

The distribution of the numerical features in the dataset can be observed from Fig. 2.10 using boxplots It has the following inferences:

- **a.** The boxplots for ***'length_of_service'*** and ***'age'*** display a moderate spread with several outliers, indicating that although most employees have a similar tenure and age range, there are notable exceptions with significantly longer service periods or higher ages.
- **b.** The ***'avg_training_score'*** boxplot reveals a relatively even distribution, suggesting consistent training performance across the workforce with few outliers, indicating the presence of some exceptionally high or low scores.

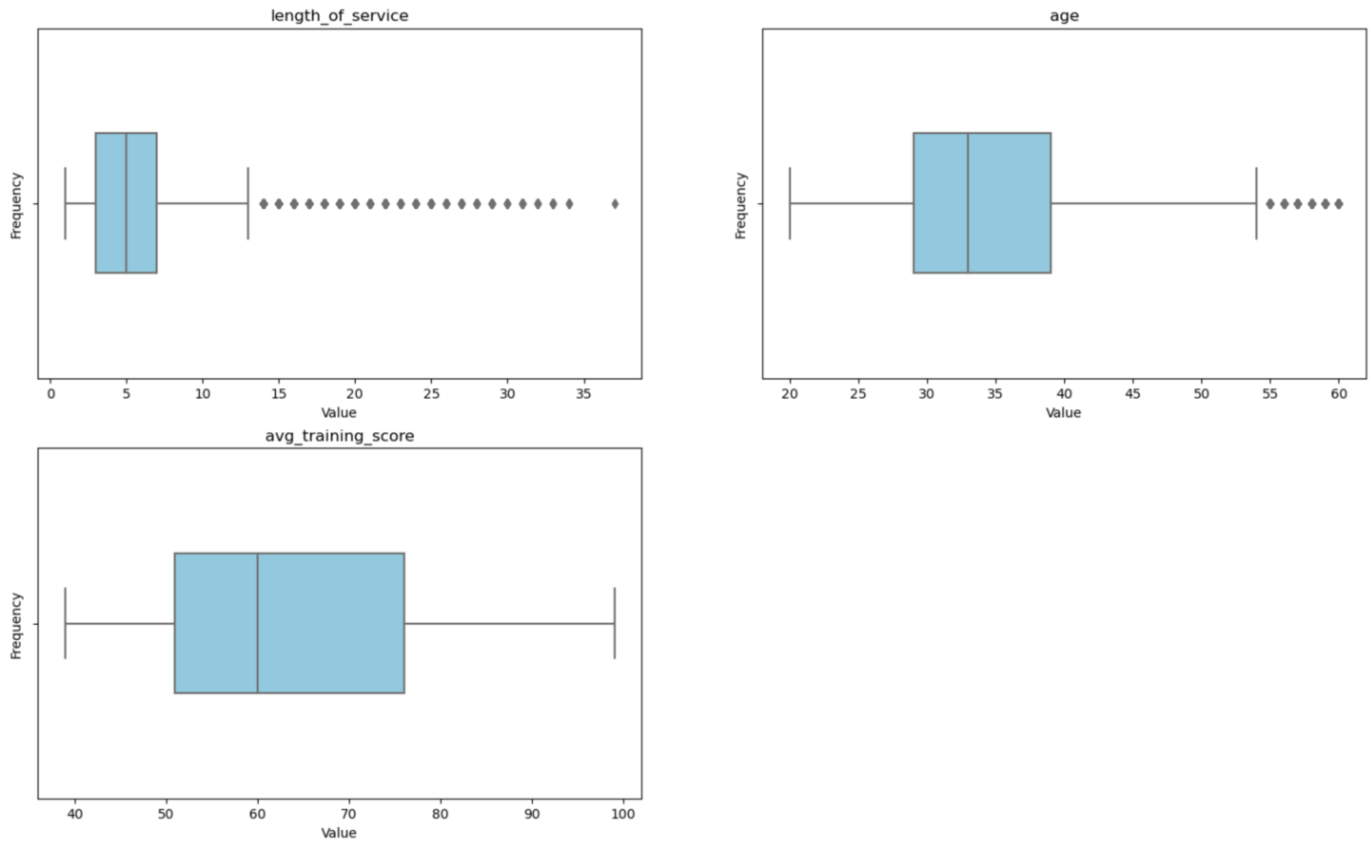

Figure 2.10. Distribution of Numerical Features

4. Distribution of Categorical Features:

From Fig. 2.11, distribution of the categorical features can be observed using count plots:

1. **Department**: The bar chart for ***'department'*** shows a higher concentration of employees in certain departments like **Sales & Marketing and Operations**, suggesting larger staffing needs or a broader scope of work in these areas.
2. **Education**: The distribution of ***education*** levels indicates that a majority of employees hold a **'Bachelor's' degree'**, followed by **'Master's & above'**, reflecting the educational qualifications typically required for roles within the organization.
3. **Gender and Recruitment Channel**: The gender bar chart reveals an **imbalance** with a predominance of **male** employees. Meanwhile, the recruitment channel graph shows that most employees are sourced or referred, highlighting the organization's preferred hiring practices.

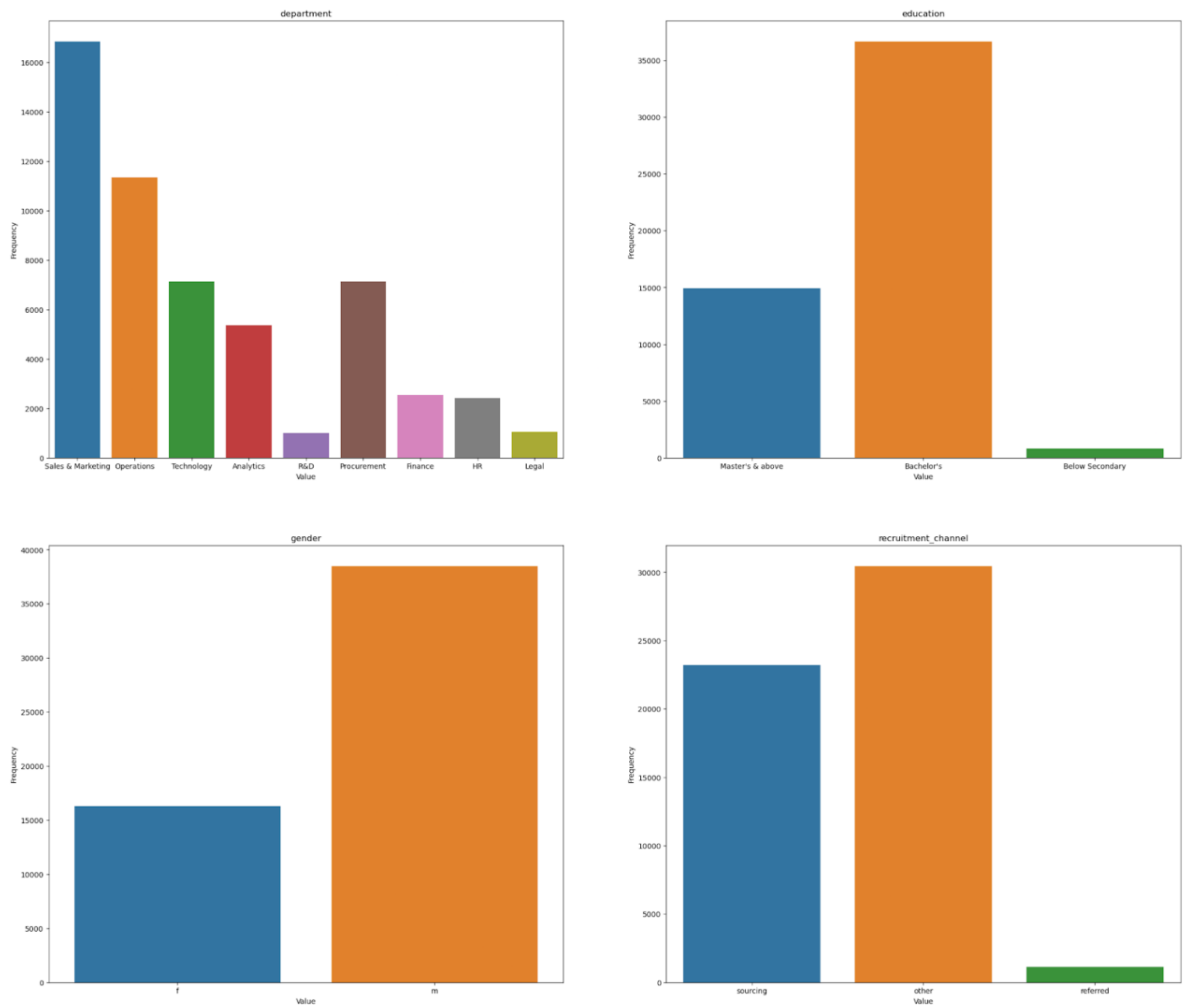

Figure 2.11. Distribution of Categorical Features

5. Distribution of Numerical Variables Based on Promotion Status:

From Fig. 2.12, distribution of the numerical features based on the promotion status can be observed:

1. **Previous Year Rating Score**: Employees who received promotions generally exhibit a higher median previous year rating compared to those who were not promoted, highlighting performance ratings as a potential influencer in promotion decisions.

2. **Age and Tenure**: The boxplots for ***'age'*** and ***'length_of_service'*** display a **broad interquartile range** among those promoted, suggesting that promotion decisions may consider factors beyond seniority, such as performance.
3. **Training Participation**: The boxplot for ***'Number of Trainings'*** indicates a relatively uniform distribution between employees who were promoted and those who were not, suggesting that merely attending a higher number of trainings does not distinctly influence promotion outcomes.

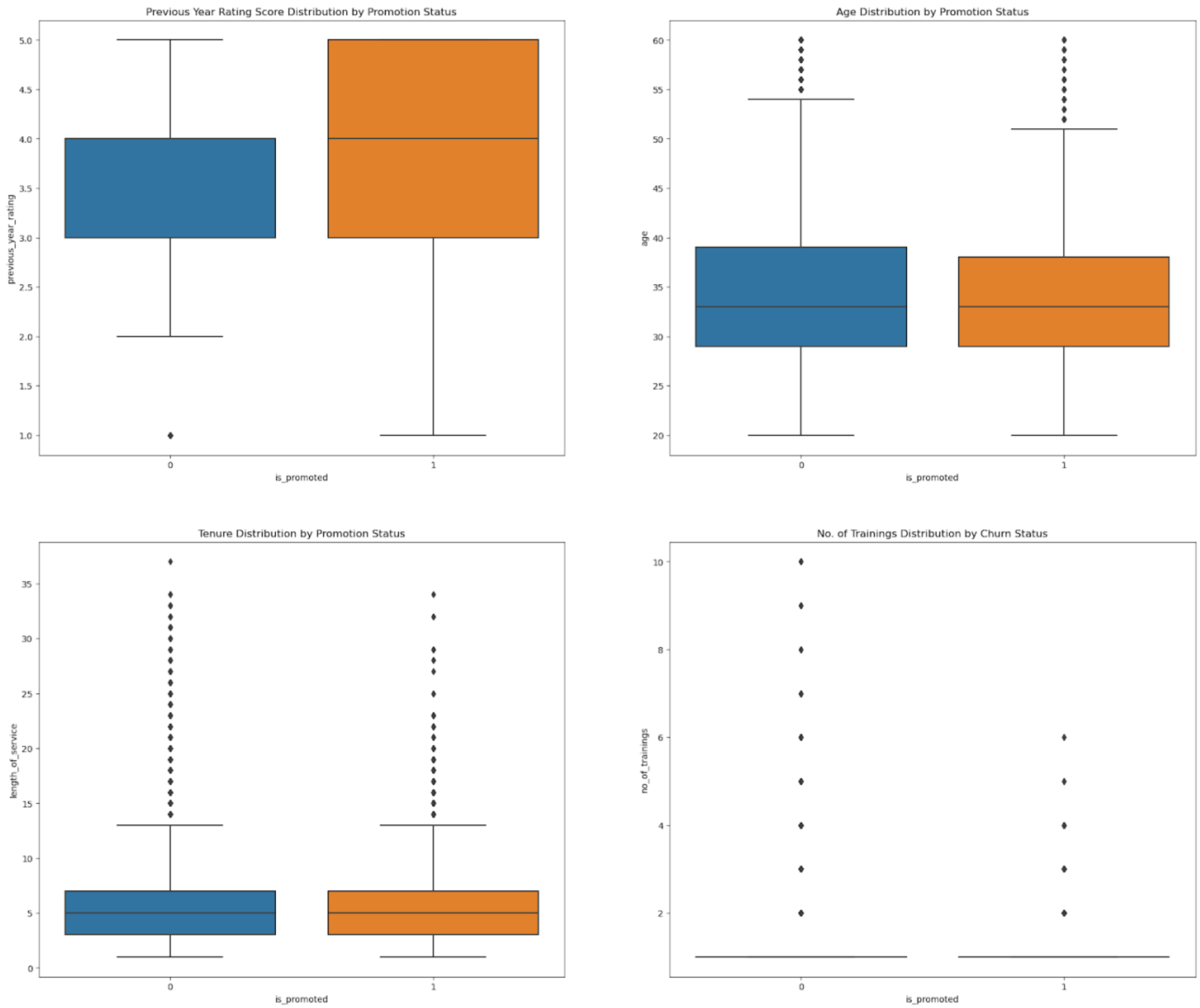

Figure 2.12. Distribution of Numerical Features on Promotion Status

6. Distribution of Training Score Based on Target Column:

The density plot from Fig. 2.13 for ***'Average Training Score'*** by ‘***promotion status’*** shows that employees who are promoted tend to have a **higher concentration** of scores in the **upper range** compared to those not promoted. This observation suggests that higher training scores could be indicative of promotion potential.

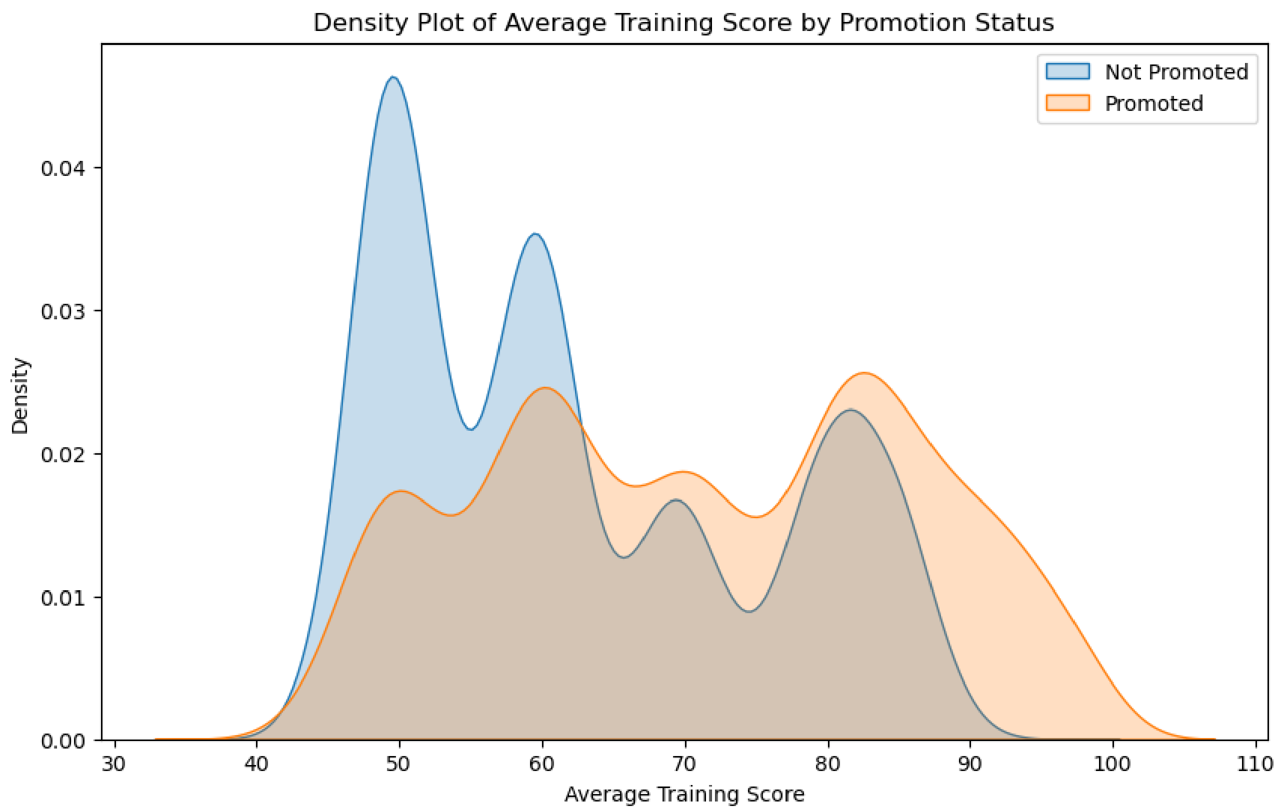

Figure 2.13. Density Plot of Average Training Score on Promotion Status

2.3.3 Feature Engineering:

1. Imputation of Previous Year Rating:

The ***'previous_year_rating'*** attribute had 4,124 missing entries. To address this, a custom imputation method was developed, leveraging the ***'length_of_service'*** of employees. The underlying assumption was that the rating should correlate with the employee's duration of service.

This ensures that the model accurately reflects the typical progression of employee ratings over time, offering a realistic and data-informed fill for missing values.

The approach was as follows:

1. For employees with no service length (new employees), it was assumed that there was no previous rating, and thus a rating of 0 was imputed.
2. For other employees, the median rating of peers with the same ***'length_of_service'*** was imputed, based on the assumption that peers with similar service durations would have comparable ratings.

2. Imputation of Education:

The ***'education'*** feature was addressed using a **mode imputation** strategy. This approach assumes that the most frequently occurring education level in the dataset is a reasonable estimate for missing data. Mode imputation is particularly effective for categorical data with a natural tendency for a common category to occur, which is often the case with education levels in a workforce.

3. Encoding Categorical Variables:

The dataset includes categorical features which are converted into a machine-readable format using **Label Encoding**, which assigns a unique integer to each category.

4. Scaling Numerical Features:

Numerical features such as ***'age'***, ***'length_of_service'***, and ***'avg_training_score'*** are standardized using a **Standard Scaler** to ensure that each feature contributes equally to the model's performance and preventing features with larger scales from disproportionately influencing the predictions.

5. Feature Interaction:

An additional feature ***'age_length_service_interaction'*** is created to capture the interaction between ***'age'*** and ***'length_of_service'***. This feature can potentially enhance the model's ability to identify complex patterns that are not evident when considering the two features separately.

6. **Correlation Analysis:**

A correlation assessment was performed among numerical variables to identify significant predictors for the target variable. Notably:

1. ***'KPIs_met >80%'*** and ***'awards_won?'*** exhibit the strongest positive correlations with the target variable, suggesting that these factors are influential in promotion decisions.
2. ***'Previous_year_rating'*** and ***'avg_training_score'*** also display positive correlations, indicating that performance and skill development are important in advancing employees.

2.4 Machine Learning Modeling

Using different models allows for assessment under various conditions, exploring how each bootstrapping method performs in estimating parameter variance, handling complexity, and applying regularization helps get a broader perspective while performing a comparative analysis.

In this project, **Linear Regression**, **Ridge Regression**, and **Random Forest Regression** are the models chosen for regression analysis the following reasons:

1. Linear Regression assumes linear relationships but struggles with outliers and multicollinearity.
2. Ridge Regression uses L2 regularization to reduce overfitting, useful when predictors outnumber observations.
3. Random Forest Regression enhances accuracy through ensemble methods, managing complex data patterns effectively without specific data relationship assumptions.

For the classification, **Logistic Regression**, **Random Forest Classifier**, and **Gradient Boosting Classifier** were chosen for the following reasons:

1. Logistic Regression is ideal for binary outcomes and offers probabilistic interpretations, however it struggles with nonlinear relationships and high-dimensional spaces without regularization.
2. Random Forest Classifier effectively handles various data types and complexities due to its robust ensemble approach, preventing overfitting.

3. Gradient Boosting Classifier builds on weak models, typically decision trees, to create a strong predictor, achieving high performance but requiring careful tuning to avoid overfitting.

Due to the imbalance in the HR Analytics dataset, **SMOTE** (Synthetic Minority Over-sampling Technique) was employed as part of an imbalanced-learn pipeline to address this issue. SMOTE works by synthesizing new examples from the minority class, thereby balancing the class distribution without losing valuable data.

This method involves creating synthetic samples rather than over-sampling with replacement, which helps provide a more robust and generalized model by broadening the decision region of the minority class. This ensures that the data remains balanced through each step of the model training process, enhancing predictive performance and accuracy.

2.5 Modeling

The parameter choices for the Bag of Little Bootstraps (BLB) and standard bootstrapping are select to balance computational efficiency and statistical robustness:

1. **Resample Size**: This was chosen as **100** to ensure faster computations while preserving the integrity of the dataset, ideal for large datasets.

2. **Number of Subsamples**: No. of Subsamples chosen as **30** to improve the statistical diversity of the samples, enhancing reliability without significantly increasing the computational burden.

Both methods use **10,000 resamples** to ensure thorough and precise sampling, enhancing the accuracy of statistical estimates such as means and confidence intervals.

This allows for rigorous comparative analysis under controlled conditions, assessing the efficiency and effectiveness of BLB against standard bootstrapping.

2.6 Metrics

In this project, several metrics were used to evaluate a model's performance on both BLB and Standard Bootstrapping. The following two metrics are used to assess the performance for both regression and classification models:

1. **CI Width**: This measures the precision of model estimates through the narrowness of confidence intervals, indicating reliability under varying data subsets.
2. **Coverage**: It assesses whether the true model parameters frequently fall within the confidence intervals, verifying the accuracy of uncertainty estimations.

For evaluating regression models, the following metrics are used along with CI Width and Coverage:

1. **Bias**: It measures the average difference between the predictions of the model and the actual values, offering insights into the model's accuracy and tendency to systematically over or underpredict across different data subsets.
2. **Mean Squared Error (MSE)**: This quantifies the average squared difference between the predicted and actual values, useful for assessing the model's overall prediction error and stability under varying bootstrapped samples.

Similarly, metrics specific to classification analysis are:

1. **Accuracy**: It provides a direct measure of the model's overall prediction correctness, valuable for evaluating performance consistency across resamples.
2. **Recall**: It focuses on the model's capability to identify all positive cases, important for assessing performance in imbalanced datasets through bootstrapping.
3. **F1 Score**: It balances precision and recall in a single metric, useful for evaluating model performance in handling class imbalances across different data samples.
4. **ROC AUC**: This evaluates the model's discriminative ability at different thresholds, offering insights into performance stability across varied training subsets in bootstrapping analysis.

3. Results

3.1 Regression Analysis:

The results for Regression analysis for each of the models can be observed in Table 3.1.

Model	Model	CI Width	Coverage	Bias	MSE
Linear Regression	BLB	2.178793	0.0.960000	0.000133	0.678931
Linear Regression	Standard Bootstrap	0.052384	-0.000009	0.157435	0.157435
Ridge Regression	BLB	1.793759	0.960000	0.000064	0.422461
Ridge Regression	Standard Bootstrap	0.039150	0.039049	0.000020	0.161780
Random Forest	BLB	1.906864	0.958333	-0.000024	0.519543
Random Forest	Standard Bootstrap	0.173052	0.226684	-0.000395	0.087697

Table 3.1. Results – Regression Analysis

Inferences:

1. **Linear Regression**: BLB exhibits wider confidence intervals and higher MSE compared to standard bootstrap but achieves significantly better coverage, indicating more robust parameter estimation despite increased variability.
2. **Ridge Regression**: Like Linear Regression, BLB with Ridge Regression shows wider confidence intervals and slightly higher MSE, yet maintains high coverage compared to very low coverage with standard bootstrap, suggesting a more balanced error minimization and true effect capture.
3. **Random Forest**: Under BLB, it also shows wider confidence intervals and higher MSE but maintains closer and higher coverage compared to standard bootstrap, implying a more realistic assessment of model uncertainty with BLB.

It can be observed that BLB consistently demonstrates wider confidence intervals and higher MSEs across all models but achieves significantly better coverage, indicating more reliable parameter estimations. This suggests that standard bootstrap might underestimate model variability, especially in complex models, and BLB offers a more accurate reflection of model performance in these cases.

3.2 Classification Analysis:

The results for Classification analysis for each of the models can be observed in Table 3.2.

Model	**Model**	**CI Width**	**Coverage**	**Accuracy**	**Recall**	**F1 Score**	**ROC AUC**
Logistic Regression	BLB	0.897853	0.108189	0.814000	0.865482	0.449435	0.911663
Logistic Regression	Standard Bootstrap	0.044550	0.061195	0.720460	0.719404	0.304776	0.807317
Random Forest Classification	BLB	0.845236	0.974000	1.000000	1.000000	1.000000	1.000000
Random Forest Classification	Standard Bootstrap	0.182471	0.084367	0.973474	0.742666	0.826654	0.976648
Gradient Boosting Classification	BLB	0.966419	0.956333	1.000000	1.000000	1.000000	1.000000
Gradient Boosting Classification	Standard Bootstrap	0.108189	0.077598	0.736823	0.894978	0.366851	0.891973

Table 3.2. Results – Classification Analysis

Inferences:

1. **Logistic Regression**: Under BLB, Logistic Regression shows a wide confidence interval, high coverage, and better performance metrics (Accuracy, Recall, F1 Score, ROC AUC) compared to standard bootstrap. This suggests BLB provides a more stable and accurate estimate of model parameters and performance.
2. **Random Forest Classification**: It achieves perfect scores for all metrics under BLB, suggesting overfitting or an excessively optimistic estimation. The standard bootstrap shows lower metrics but still substantial ROC AUC, indicating a possibly more realistic evaluation but with very low coverage, hinting at underestimation of parameter variability.
3. **Gradient Boosting Classification**: Like Random Forest, Gradient Boosting scores perfectly under BLB across all metrics, potentially reflecting overfitting. The standard bootstrap, while still performing well, shows lower scores and very low coverage, again suggesting a potential underestimation of true variability.

BLB consistently shows wider confidence intervals and higher coverage compared to standard bootstrap across all models, suggesting more robust and reliable parameter estimation. However, the perfect scores under BLB for ensemble methods like Random Forest and Gradient Boosting may indicate overfitting. This highlights the importance of balancing accuracy with realistic estimations in model evaluation, especially with complex classifiers.

4. Results

This project highlights the critical importance of selecting the appropriate methodological approach tailored to the characteristics of the datasets and available computational resources. Some of the key-takeaways from this project are:

Bag of Little Bootstraps (BLB) vs. Standard Bootstrap:

The project utilized both the Bag of Little Bootstraps (BLB) and standard bootstrap methods to evaluate the stability and reliability of our regression and classification models. Particularly, the BLB method proved exceptionally effective for large datasets, offering significant advantages in computational efficiency and scalability.

BLB's Advantages: By resampling smaller subsets of data, the BLB method significantly reduces computational demands while still delivering accurate uncertainty estimates. This feature is especially beneficial for large datasets, where traditional bootstrap methods might not be practical due to their extensive computational requirements.

BLB Limitations: Despite its advantages, the BLB method can introduce bias if the subsets are not representative of the whole, potentially leading to less accurate estimates compared to the full data bootstrap. Additionally, the choice of parameters for defining the size and number of subsets can significantly impact the results, requiring careful consideration to ensure robust outcomes.

5. Acknowledgments

Given the computationally intensive nature of running 10,000 resamples, we leveraged our university's iLab resources. The HPC capabilities of iLabs provided the necessary computational power, allowing us to conduct extensive bootstrap analysis without the constraints that would typically limit such a study and significantly reducing the computation time.

6. References

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(4):795–816