Bootstrap Analysis - Bag of Little Bootstraps

Group - 9

***Rohit** Kulkarni (**rk1169**)*
*Naga Satya **Dheeraj** Anumala (**na945**)*
***Krit** Gupta (**ksg124**)*
*Sai **Adarsh** Kasula (**sk2837**)*

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

GitHub: https://github.com/rohitkulkarni08/enhancing-predictive-modeling-using-bootstrapping

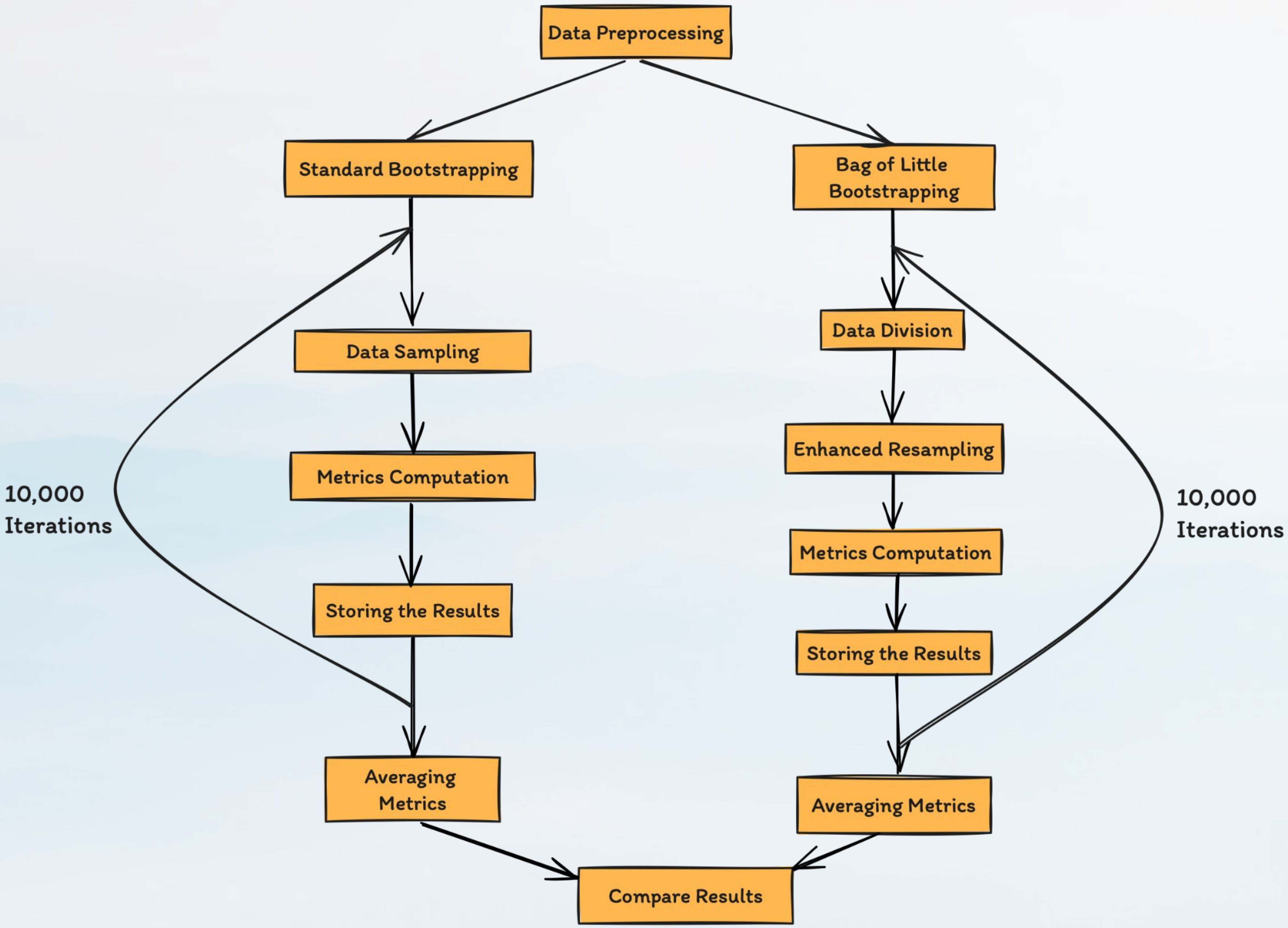
Bootstrap Analysis - Bag of Little Bootstraps

In this project, we explore the **Bag of Little Bootstraps (BLB)** method, a scalable bootstrap technique designed for assessing the quality of estimators in massive datasets.

It combines the advantages of traditional bootstrap and subsampling approaches, offering a computationally efficient means of estimating uncertainties in large-scale data analysis.

We perform a comparative analysis of Bag of Little and Standard Bootstrapping on Predictive Modeling using large-scale datasets.

Leveraged **Rutgers iLabs** to run bootstrapping calculations



Datasets

Two distinct datasets are used in this project, both taken from **Kaggle** for Regression and Classification Analysis

Regression Analysis:

Dataset: Zomato.

Target Variable: Average Order Cost for Two people

Features: Contains information on various restaurants such as location, cuisine types, ratings, and the typical cost for a meal for two.

Classification Analysis Dataset:

Dataset: HR Analytics.

Target Variable: Employee promotion status

Features: Performance ratings, years of experience, goals met, department, age, and training scores.

Exploratory Data Analysis and Feature Engineering

1. ***Data Cleaning and Transformation***:

To enhance the dataset's, following transformations were performed:

1. ***Missing Cost Data Handling***
2. ***Scaling Numerical Variables***
3. ***Removing Unnecessary columns***
4. ***Encoding Cat Variables***

2. ***Vectorizing cuisine data:***

Cuisines feature **vectorized** using **Word2Vec** to create numerical vectors, making it suitable for machine learning analysis

3. ***Imputation of Missing Ratings:***

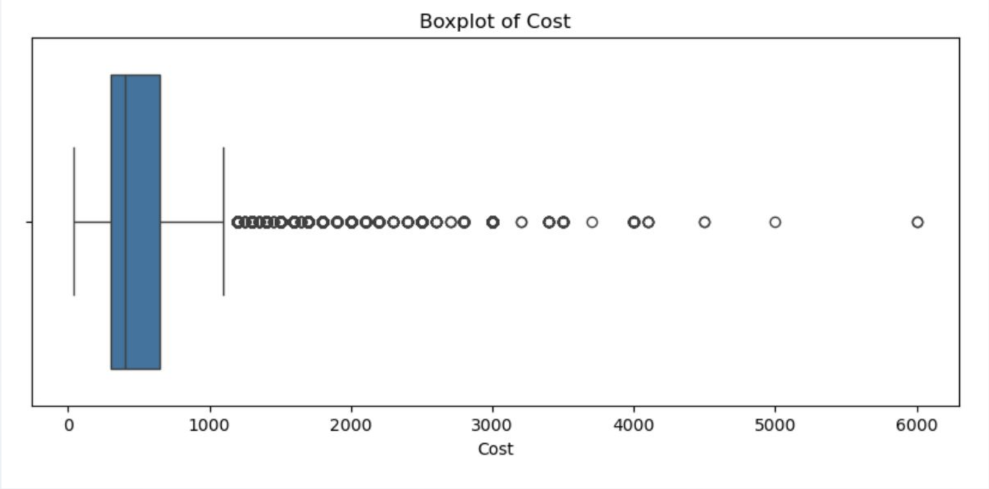
MICE with **Linear Regression** was employed to estimate missing rate entries, improving dataset completeness.

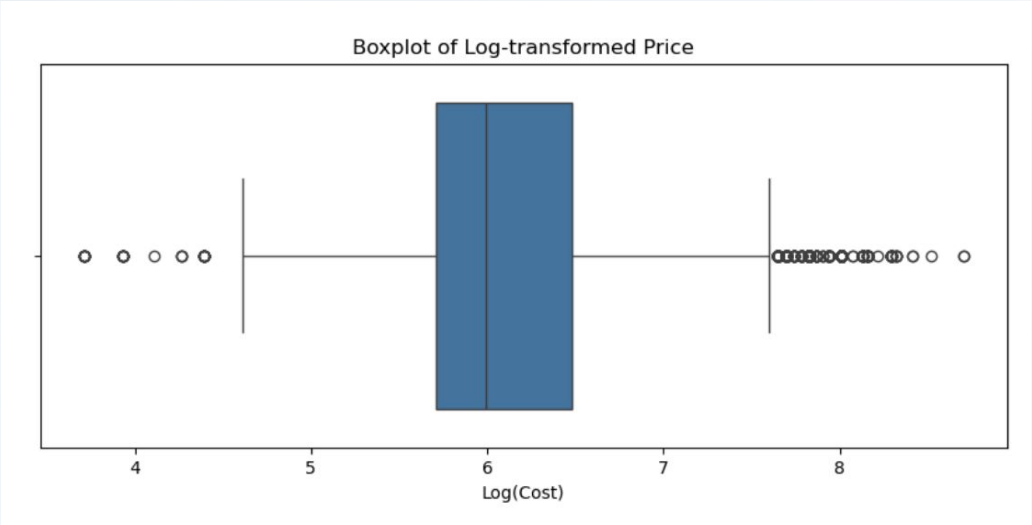
4. ***Cost Data Normalization***

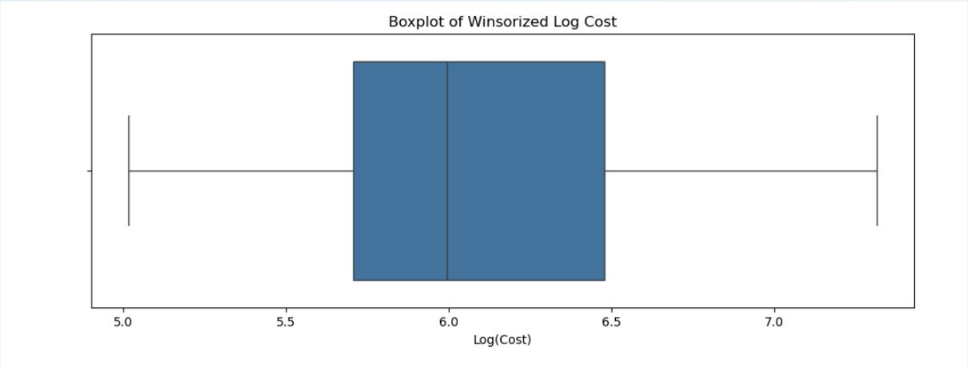
There are a lot of outliers for the average cost with and a notable **skew** of **2.60198**

Log transformation was applied which **reduced skewness**. however, there were still a **few outliers** present in the average cost.

Finally, ***Winsorization*** was applied to the log transformed cost in order to scale the extreme values and remove the remaining outliers, enhancing data uniformity.







Exploratory Data Analysis and Feature Engineering

1. ***Imputing Previous Year Rating:***

Missing entries are treated using a custom imputation method was developed using ***Length of Service***

New employees (employees with no service length) received a rating imputation of 0, assuming no prior rating.

For the remaining employees, the **median** rating of peers with the same ***Length of Service*** was imputed, reflecting comparable peer ratings over similar durations.

2. ***Imputing of Education:***

Missing data was handled using ***mode imputation***, This method is effective where certain categories dominate the dataset

3. ***Categorical Encoding***:

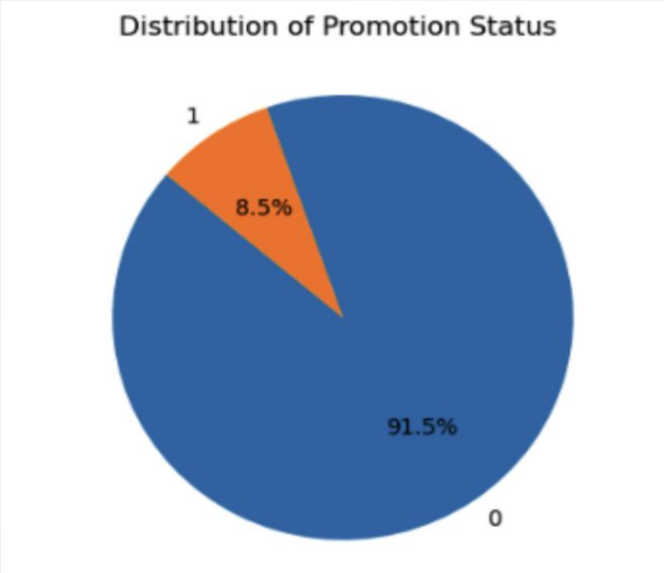
Categorical features are converted into a machine readable format using ***Label Encoding***,

4. ***Scaling Numerical Features:***

Numerical features were standardized using a ***Standard Scaler*** to ensure equal contribution to the model and prevent features with larger scales from disproportionately influencing predictions.

5. ***Target Variable Analysis:***

There is a significant imbalance in the ***Promotion Status,*** which is dealt using **SMOTE** while building the classification models



Modeling

In this project, **Linear Regression, Ridge Regression, and Random Forest Regression** are the models chosen for regression analysis; and **Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier** were chosen for classification analysis.

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

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

2. **No. of Subsamples** chosen as ***30*** to improve the statistical diversity and enhancing reliability without impacting the computational power.

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

Metrics:

The following metrics are used to assess the performance for **both** regression and classification: 1. ***CI Width*** 2. ***Coverage***

For evaluating **regression models**, the following metrics are used along with **CI Width** and **Coverage**: 1. ***Bias*** 2. ***Mean Squared Error***

Similarly, metrics specific to **classification analysis** are: 1. ***Accuracy*** 2. ***Recall*** 3. ***F1-Score*** 4. ***ROC-AUC***

Results: Regression Analysis

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

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:

Ridge Regression shows **wider CI** 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 Regression:

Under BLB, it also shows **wider confidence intervals** and **higher MSE** but maintains **closer** and **higher coverage** with standard bootstrap, implying a more realistic assessment of model uncertainty with BLB.

Results: Classification Analysis

Model	Model	CI Width	Coverage	Accuracy	Recall	F1 Score	ROC AUC
Logistic Regression	Bag of Little Bootstraps	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	Bag of Little Bootstraps	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	Bag of Little Bootstraps	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

1. Logistic Regression:

Under BLB, it shows a wide confidence interval, high coverage, and better performance metrics 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. Traditional 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, it scores perfectly under BLB, potentially reflecting overfitting. Standard bootstrap, shows lower scores and very low coverage, suggesting a potential underestimation of true variability.