Project : Bank Marketing
Name : Apoorv Ajay Thite
University : The Pennsylvania State University
Email : apoorvthite21@gmail.com
Country : United States of America
Specialisation : Data Science
Internship Batch : LISUM34
Date : 25th July, 2024

Outliers and Skewness Handling and Associated Data Operations

Problem Description

ABC Bank aims to predict whether a customer will subscribe to their term deposit product based on their past interactions with the bank or other financial institutions. The data is sourced from direct marketing campaigns conducted by a Portuguese banking institution, primarily through phone calls. The classification goal is to predict if a client will subscribe ('yes') or not ('no') to a term deposit, helping the bank to focus their marketing efforts efficiently and save resources.

To ensure the robustness and accuracy of the predictive models for ABC Bank's marketing campaign, it was crucial to address data quality issues such as outliers and skewness. I employed the Interquartile Range (IQR) method to handle outliers and the log transformation method to manage skewness, ensuring that the dataset was in optimal condition for modelling.

Handling Outliers with the IQR Method

Outliers can significantly distort the performance of machine learning models, leading to inaccurate predictions. To mitigate this, I used the IQR method to detect and remove outliers from the dataset. The IQR method involves calculating the first quartile (Q1) and the third quartile (Q3) of the data, with the IQR being the difference between these two values (IQR = Q3 - Q1). I then defined the lower bound as Q1 - 1.5 * IQR and the upper bound as Q3 + 1.5 * IQR. Data points falling outside these bounds were considered outliers and were removed from the dataset. This method is effective because it is not influenced by extreme values and provides a robust way to identify and eliminate anomalies in the data.

Addressing Skewness with Log Transformation

Skewness in data can affect the distribution and subsequently the performance of machine learning models. To normalize the distribution, I applied the log transformation method to skewed numerical columns. Log transformation involves taking the natural logarithm of each data point, which reduces the impact of larger values and helps in achieving a more

symmetric distribution. This transformation is particularly useful for positively skewed data, as it compresses the range of the data, bringing extreme values closer to the centre. By normalizing the data distribution, the log transformation improves the model's ability to learn and generalize from the data.

Handling Missing Values

Upon examining the dataset, I found that there were no missing values. This eliminated the need for imputation or other techniques to handle missing data, simplifying the preprocessing steps. Having a complete dataset allowed me to focus more on addressing outliers and skewness, ensuring that the data was clean and well-prepared for the subsequent modelling stages.

By meticulously addressing these data quality issues, I ensured that the predictive models developed for ABC Bank were based on a solid foundation, leading to more accurate and reliable predictions. This comprehensive data preprocessing approach is critical for the success of any data-driven project, particularly in the context of predicting customer behavior for marketing campaigns.