RADemics

Comprehensive Data Preprocessing and Feature Engineering for Optimized Machine Learning Models

Rama Nandan Tripathi, Dileep Kumar

DR. RAM MANOHAR LOHIA AVADH UNIVERSITY, DR. RAM MANOHAR LOHIA AVADH UNIVERSITY

Comprehensive Data Preprocessing and Feature Engineering for Optimized Machine Learning Models

Rama Nandan Tripathi, Assistant Professor MCA, Dr. Ram Manohar Lohia Avadh University Ayodhya U.P. – 224001. sonu.ramanandan@gmail.com

Dileep Kumar, Designation- Assistant Professor MCA, Dr. Ram Manohar Lohia Avadh University Ayodhya U.P. – 224001. dileep_k_2000@yahoo.com

Abstract

In machine learning, the quality of feature engineering and data preparation has a major impact on how effective predictive models are. This chapter offers a thorough analysis of sophisticated feature engineering and data pretreatment methods, emphasizing their vital importance in machine learning model optimization. Emphasis was placed on data cleaning methods, including automated tools for handling missing values and outlier detection, which are essential for ensuring data integrity. Additionally, the chapter explores sophisticated feature engineering practices that enhance model performance, such as dimensionality reduction, feature selection, and transformation techniques. The interplay between data quality and model accuracy was critically analyzed, highlighting the importance of robust preprocessing strategies in achieving reliable and effective machine learning outcomes. Key advancements in automated data cleaning and feature engineering are discussed, alongside their practical implications for real-world applications. This chapter serves as a crucial resource for researchers and practitioners seeking to enhance their understanding of data preprocessing and feature engineering to improve machine learning model performance.

Keywords: Data Preprocessing, Feature Engineering, Automated Data Cleaning, Missing Values Imputation, Outlier Detection, Machine Learning Models.

Introduction

The quality of data preprocessing and feature engineering was critical for creating accurate and dependable prediction models in the quickly emerging field of machine learning [1-3]. Efficient data preparation was a sequence of actions intended to get raw data ready for analysis, such as data integration, transformation, and cleaning [4,5]. The quality of the data that machine learning models are fed directly affects their performance and accuracy [6-8]. The necessity for sophisticated preparation procedures increases with the amount and complexity of datasets. The objective of this chapter was to present a thorough review of state-of-the-art techniques for feature engineering and data preparation, emphasizing their importance for improving model efficacy and guaranteeing reliable data analysis.

An essential component of data preparation was data cleaning, which deals with problems including missing values, outliers, and inconsistencies that can negatively impact model performance. This procedure has been transformed by automated data cleaning technologies, which make it easier to find and fix problems with data quality [9]. In order to properly prepare datasets for analysis, methods like mean imputation, multiple imputation, and advanced outlier identification algorithms are essential. In addition to increasing productivity, the use of automated technologies guarantees a greater level of accuracy while managing complicated and substantial amounts of data [10-12]. This chapter delves further into various instruments and techniques, offering perspectives on their usefulness and implementation.

In order to improve the performance of machine learning models, feature engineering entails the generation and selection of pertinent features from raw data [13]. Techniques including feature selection, data transformation, and dimensionality reduction are used in this procedure [14]. High-dimensional data management was aided by dimensionality reduction techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), which reduce the amount of features while maintaining crucial information [15]. By identifying the most useful features, feature selection techniques like LASSO and recursive feature removal help reduce overfitting and improve the interpretability of the model [16]. This chapter explores these sophisticated methods and shows how affect the performance of the model.