

Chapter 1


Feature Engineering for Various Data Types in Data Science

Nilesh Kumar Sahu

 <https://orcid.org/0000-0003-1675-7270>

Birla Institute of Technology, Mesra, India

Manorama Patnaik

 <https://orcid.org/0000-0003-4035-2468>

Birla Institute of Technology, Mesra, India

Itu Snigdh

Birla Institute of Technology, Mesra, India

ABSTRACT

The precision of any machine learning algorithm depends on the data set, its suitability, and its volume. Therefore, data and its characteristics have currently become the predominant components of any predictive or precision-based domain like machine learning. Feature engineering refers to the process of changing and preparing this input data so that it is ready for training machine learning models. Several features such as categorical, numerical, mixed, date, and time are to be considered for feature extraction in feature engineering. Datasets containing characteristics such as cardinality, missing data, and rare labels for categorical features, distribution, outliers, and magnitude are currently considered as features. This chapter discusses various data types and their techniques for applying to feature engineering. This chapter also focuses on the implementation of various data techniques for feature extraction.

INTRODUCTION

The process of changing and preparing input data trained to be ready for machine learning models is called Feature Engineering. Features such as Categorical, Numerical, Mixed and date and time are to be considered for feature extraction in feature engineering. Datasets containing features such as cardinality, missing data and rare labels for categorical features, distribution, outliers and magnitude are being

DOI: 10.4018/978-1-7998-6659-6.ch001

considered as features. This chapter discusses about various data types and their techniques applied in feature engineering. This chapter also focuses on implementation of various data techniques for feature extraction.

Feature Engineering is a process of transforming the raw data from one form to another such that it can be represented in a better way (Kuhn & Johnson, 2019). In a short we can say creating new features from the existing list of features such that it will help in the improvement of learning model performance (Ruder et.al, 2019). Feature engineering became out of the desire to change linear regression inputs that are not typically distributed (Bengio et.al., 2013). Such change can be useful for linear regression. The original work by George Box and David Cox in 1964 presented a technique for figuring out which of a few force capacities may be a valuable change for the result of Linear Regression (Box, 1964). This is now known as the Box-Cox change (Tommaso, 2011). Linear regression isn't the main machine learning model that can benefit from highlight building and different changes. In 1999, it was shown that element building could improve the presentation of rules learning for text classification (Heaton, 2016).

Feature engineering is the assignment of improving prescient demonstrating execution on a dataset by changing its component space (Coates, et.al, 2011). Existing ways to deal with mechanize this procedure depends on either changed component space investigation through assessment guided hunt, or unequivocal extension of datasets with every single changed element followed by include determination (Scott & Matwin, 1999). Such methodologies acquire high computational expenses in runtime and additionally memory. A novel procedure for learning Feature Engineering (LFE) is presented, for robotizing feature building in classification errands which depends on learning the adequacy of applying a change (e.g., number-crunching or total administrators) on numerical highlights, from past component designing encounters. Given another dataset (Nargesian, et.al, 2017), LFE prescribes resource of helpful changes to be applied on highlights without depending on model assessment or express component development and determination (Krasanakis, et.al, 2018). Utilizing an assortment of datasets, we train a lot of neural systems, which target anticipating the change that impacts classification execution decidedly (Jiang, et.al, 2008).

This chapter is presenting section 1 as Introduction, section 2 is discussing on data types in data Science, section 3 is focussing on Different Techniques of applying Feature Engineering, section 4 illustrates Different Techniques of applying Feature Engineering, section 5 presents conclusion and section 6 describes the future work .

Goal of Feature Engineering in Data Science

The main goal of feature engineering is to remove unwanted features from the given raw dataset while keeping the main and important features, which will help us to derive some useful and important information's (Khurana, et.al, 2016).

Why Feature Engineering Is Needed?

Feature Engineering is needed for increasing the accuracy of the learned model such that the model which is being trained on the given data can also achieve better accuracy on the unseen data too (Weiss, et.al, 2016) .

According to the latest survey in Forbes, Data Scientist spend around 75% of their time, just on data preparation (Dong, et.al, 2018).

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/feature-engineering-for-various-data-types-in-data-science/268746

Related Content

Overview of Big Data With Machine Learning Approach

(2021). *Machine Learning in Cancer Research With Applications in Colon Cancer and Big Data Analysis* (pp. 160-189).

www.irma-international.org/chapter/overview-of-big-data-with-machine-learning-approach/277022

Quorum Sensing Digital Simulations for the Emergence of Scalable and Cooperative Artificial Networks

Nedjma Djezzar, Iñaki Fernández Pérez, Noureddine Djediand Yves Duthen (2019). *International Journal of Artificial Intelligence and Machine Learning* (pp. 13-34).

www.irma-international.org/article/quorum-sensing-digital-simulations-for-the-emergence-of-scalable-and-cooperative-artificial-networks/233888

The Dark Side of AI: Adversarial Attacks and Defenses in Deep Learning

Uddalak Mitraand Shafiq UI Rehman (2025). *Challenges and Solutions for Cybersecurity and Adversarial Machine Learning* (pp. 1-36).

www.irma-international.org/chapter/the-dark-side-of-ai/382256

Rule Extraction in Trained Feedforward Deep Neural Networks: Integrating Cosine Similarity and Logic for Explainability

Pablo Ariel Negroand Claudia Pons (2024). *International Journal of Artificial Intelligence and Machine Learning* (pp. 1-22).

www.irma-international.org/article/rule-extraction-in-trained-feedforward-deep-neural-networks/347988

Machine Learning in the Catering Industry

Lanting Yang, Haoyu Liuand Pi-Ying Yen (2023). *Encyclopedia of Data Science and Machine Learning* (pp. 277-285).

www.irma-international.org/chapter/machine-learning-in-the-catering-industry/317452