International Journal of Artificial Intelligence, Computer Science, Management and Technology
Volume 2 Issue 3 JUL – SEP 2025 | International Peer Reviewed & Refereed Journal

Feature Engineering and Selection Strategies for Improving Machine Learning Accuracy

Meena Kumari
Independent Researcher

Abstract

A significant contribution to the enhancement of the precision and dependability of machine learning models is made by the processes of feature engineering and feature selection. When dealing with complicated and high-dimensional datasets, the quality of the input features frequently has a higher impact on the performance of the model than the algorithm that is selected itself. Feature engineering is the process of transforming raw data into informative representations, whereas feature selection is the process of identifying the variables that are most significant, hence eliminating noise and repeated information. These techniques for feature engineering include normalization, encoding, feature building, and domain-driven transformations. Additionally, these techniques include feature selection strategies such as filter, wrapper, and embedding methods. This article will discuss how these approaches improve model generalization, decrease overfitting, and enhance computing efficiency across a variety of machine learning problems. An analysis of the impact of feature engineering and selection on both traditional and advanced learning models is presented, along with comparative observations for each.

Keywords: Feature Engineering, Feature Selection, Machine Learning Accuracy, Dimensionality Reduction

Introduction

When it comes to producing accurate and trustworthy predictions, machine learning models are strongly dependent on the quality and relevance of the data that they receive as input. Despite the fact that improvements in algorithmic and computational capacity have led to greater model performance, the effectiveness of a machine learning system is frequently determined more by the degree to which the input data accurately describe the underlying problem. Therefore, feature engineering and feature selection play a major role in the process of translating raw data into meaningful inputs that improve the efficiency of learning and the accuracy of predictions. The process of developing, manipulating, and encoding variables in

order to more effectively capture patterns within data is referred to as feature engineering. The normalization process, the handling of missing values, the encoding of categorical variables, and the construction of new features based on domain knowledge are all possible steps in this process. Models are able to learn relevant correlations more efficiently and lessen the influence of noise and data inconsistencies when they consist of features that have been well-engineered. When it comes to feature selection, on the other hand, the primary objective is to find and keep the features that are the most informative while deleting those that are repetitive or irrelevant. It is common for high-dimensional datasets to include variables that are not necessary, which can lead to an increase in computational complexity and an increase in the likelihood of overfitting. Better generalization, faster training, and enhanced interpretability are all outcomes that can be achieved by models through the selection of an ideal subset of features. enhancements to the accuracy of machine learning can be achieved through feature engineering and selection procedures. The strategies that are routinely utilized, the practical benefits that they offer, and the impact that they have on various types of learning models are all investigated. The purpose of this research is to provide insights that will enable the creation of machine learning systems that are efficient, accurate, and scalable. This will be accomplished by highlighting best practices concerning feature optimization.

Role of Feature Engineering in Machine Learning

One of the most important aspects of the success of machine learning models is feature engineering, which is the process of transforming raw data into features that are useful and informative. Frequently, raw datasets are characterized by the presence of noise, the absence of values, or formats that are not directly advantageous for learning algorithms. Models are able to more accurately capture underlying patterns and relationships when the data is cleaned, converted, and arranged in a manner that is accomplished through the process of feature engineering.

Having features that are well constructed can help increase the accuracy, stability, and generalization of a model. Normalization, scaling, encoding categorical variables, and handling missing values are some of the techniques that can be utilized to guarantee that the data that is input is consistent and comparable. A great number of instances include the creation of new features by the combination of pre-existing variables or the application of domain knowledge, which can reveal hidden trends that the model would not have observed otherwise.

Additionally, feature engineering lessens the complexity of the model and the amount of computational resources required. In order to improve the efficiency of learning and reduce the likelihood of overfitting, models can be improved by removing information that is irrelevant or redundant and putting greater emphasis on the features of the data that are the most informative. When dealing with high-dimensional datasets, where an excessive number of features might have a negative impact on performance, this is of crucial importance.

Impact of Feature Engineering on Model Performance

When it comes to the performance of machine learning models, feature engineering has a direct and significant impact on the overall performance. Better prediction accuracy and resilience can be achieved through the utilization of high-quality features, which make it possible for models to learn relevant patterns more efficiently. When raw data is processed and formatted in the appropriate manner, models are able to more accurately capture the relationships between variables, which in turn reduces errors and improves overall performance. The elimination of noise and the decrease of data inconsistency are two of the most important effects that feature engineering has. It is possible to ensure that features contribute in an equitable manner to the learning process by employing techniques like as normalization, scaling, and handling missing values. When it comes to methods that are sensitive to the magnitude or distribution of features, such as distance-based and gradient-based models, this is of utmost importance. Because of this, designed characteristics frequently result in a more stable training environment and a faster speed of convergence. Additionally, feature engineering enhances model generalization by reducing the amount of overfitting that occurs. Through the elimination of features that are not relevant and the construction of meaningful representations, models concentrate on the information that is most important rather than learning to memorize random fluctuations in the data. Because of this, performance on data that has not yet been seen improves, which is essential for deployment in the real world. Not only does efficient feature engineering improve accuracy, but it also improves computing efficiency and allows for greater interpretability. Models that are trained on features that have been carefully built need fewer resources, generate predictions that are more reliable, and provide greater insights into the operations of decision-making processes. In light of this, feature engineering continues to be an essential component in the process of developing machine learning systems that are both high-performing and scalable.

Evaluation Metrics and Experimental Design

When it comes to determining how effective feature engineering and selection procedures are in machine learning, evaluation metrics and experimental design are absolutely necessary components. A well-structured experimental design guarantees that performance improvements are attributable to feature optimization rather than random variation or data leakage. Appropriate metrics provide a valid assessment of how well a model performs, and a well-structured experimental design ensures that this information is used.

Depending on the kind of machine learning activity being performed, common evaluation metrics can be different. Accuracy, precision, recall, F1-score, and area under the ROC curve are some of the most common statistical measures that are utilized to evaluate prediction performance while dealing with classification issues. A number of metrics, including mean absolute error, mean squared error, and R-squared, are frequently utilized in assignments involving regression. The contribution of engineered and selected features to the outputs of the model can be quantified with the use of these measures.

The process of separating the dataset into training, validation, and testing sets is a common component of experimental design. This is done in order to assess the generality of the model. The reduction of bias and variation in performance estimation is frequently accomplished through the utilization of methods such as cross-validation. The stages of feature engineering and selection are only applied to the training data in order to prevent information leaking. This helps to ensure that the assessment results continue to be objective and credible.

In order to quantify the variations in performance, it is usual practice to run comparative experiments by training models with and without feature optimization. It is also possible to utilize statistical validation methods in order to verify the significance of the improvements that have been seen. A structured framework is provided by the combination of carefully selected assessment metrics and a rigorous experimental design. This framework allows for an understanding of how feature engineering and selection contribute to enhanced machine learning accuracy and dependability.

Conclusion

When it comes to improving the accuracy, efficiency, and reliability of machine learning models, feature engineering and feature selection are essential components. According to the findings of this study, the selection of a learning algorithm is not the only factor that can have an impact on the performance of a model; well-designed features can also have a significant

impact. The capacity of models to better capture underlying patterns and reduce the influence of noise and redundancy can be improved by the transformation of raw data into meaningful representations and the selection of the variables that are most significant. The successful engineering of features improves the generalization of models, decreases the amount of overfitting, and increases the efficiency of computational processes. When applied to high-dimensional datasets, feature selection is particularly useful since it streamlines the learning process and improves interpretability. This further enhances the performance of the learned model. These methodologies, when combined, contribute to the development of machine learning systems that are more robust and scalable. The results put an emphasis on the significance of including approaches for feature optimization into the machine learning process. In order to further improve model performance in contexts that are both complicated and constantly changing, future research should concentrate on automated feature engineering, domain-aware feature building, and adaptive selection approaches.

Bibliography

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, 249–268.

Van der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10, 66–71.

Liu, H., & Motoda, H. (2007). *Computational Methods of Feature Selection*. Chapman & Hall/CRC.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.