Nat. Volatiles & Essent. Oils, 2021;8(1):194-200

NVEO
Natural Volatiles & Essential Oils

Techniques For Feature Engineering To Improve Ml Model Accuracy

Naresh Babu Kilaru[1*], Sai Krishna Manohar Cheemakurthi[2]

[1*]Independent Researcher, nareshkv20@gmail.com
[2]Independent Researcher, saikrishnamanohar@gmail.com

***Corresponding Author:** Naresh Babu Kilaru
*Independent Researcher, nareshkv20@gmail.com

Abstract

In this paper, the author sought to understand the effects of feature engineering on the enhancements of the models learned by a machine learning algorithm. Feature engineering takes the raw data and prepares them for model inputs, increasing the model's effectiveness. Using different features on different datasets, this study assesses the performance of techniques like feature selection, feature extraction, feature scaling, feature engineering, and feature encoding. Using filter, wrapper, and embedded methods, we determine suitable features that describe a specific problem or situation well, while extraction methods such as PCA and autoencoders minimize feature dimensionality. Scaling techniques help normalize and scale the data, and the encoding methods assist in translating a categorical variable to a numerical value. As can be seen from the results, there are substantial enhancements in model performance, stability, and training time. Examples in finance, healthcare, and e-commerce are highlighted to show how these approaches solve diverse problems, such as detecting fraud, predicting diseases, or segmenting customers. There are also examples of feature selection and evaluation problems and their solutions discussed in the paper, which include issues with dimensionality and multicollinearity. In this respect, the study aims to discuss these challenges and recommend how feature engineering can be integrated to improve model performance and interpretability in real-world cases.

Keywords: Feature Engineering, Machine Learning, Model Accuracy, Feature Selection, Feature Extraction, Scaling, Encoding, High Dimensionality, Multicollinearity, Automated Tools, PCA, LASSO, Cross-Validation, Data Preprocessing, Domain Knowledge.

Introduction

Feature engineering is another feature selection or construction process in machine learning. It involves preprocessing raw data to make it more suitable for the problem under consideration in predictive models and make the models more interpretable [3]. Feature engineering is beneficial since it brings about a better understanding of the model, faster training and increases the reliability of the predictions, making it a measure of the efficiency of the machine learning solutions [2]. Therefore, this paper has presented a few experiments to compare primary feature engineering techniques, such as selection, extraction scaling, and encoding of various classifications and regression models on various datasets [3].

According to the techniques mentioned above, the research is intended to establish and exemplify their applicability and functionality in various sectors such as finance, health, and technology, mainly e-commerce [4]. Moreover, examples indicate how these methods help solve concerns about specific areas, fraud management, assessment of risks of the diseases' emergence, and identification of ordinary clients [5]. Moreover, this paper also provides possible solutions and recommendations on some problems that may arise during feature engineering, such as high dimensionality problems regarding feature vectors and the

entire dataset, multicollinearity and data sparsity issues [6]. Consequently, the design of this proposed study is to offer practical solutions aimed at improving the performance and utility of feature engineering as an essential step in advancing the field of machine learning [7].

Simulation Reports
Methodology

The simulations used datasets from different areas, such as credit card fraud, electric load and stock price movement prediction. The models used during the simulations comprised decision trees, random forests, and deep learning architectures to evaluate the effect of feature engineering on the models. A dataset was taken from Kaggle for the credit card fraud detection task. It contains the anonymized transactions made by credit cards, the main target of which is to detect fraudulent or genuine transactions as classes. In this case, since the data is imbalanced, feature engineering turned out to be very crucial [4]. The electric load forecasting task used time series data from an energy provider, which regarded hourly load measurements to forecast future demand based on updated features [7]. Last, real-time stock price movement prediction has incorporated opening price, closing price, volume, and external economic indices to enhance the model with feature engineering techniques [10].

Feature Engineering Steps Implemented
Respective feature engineering methodologies were implemented on every dataset to fine-tune the model. Correlational filter techniques were used for feature selection, which involved the elimination of redundant features while at the same time reducing the complexity of the model [6]. Other wrapper methods, such as Recursive Feature Elimination (RFE), were used to determine the best features to enhance the model's forecast capability[8]. Some embedded methods, especially the LASSO regression, also included the feature selection mechanism during the modelling process, automatically selecting the most appropriate features [6]. Aggregate feature extraction methods like PCA, which perform dimensionality reduction, effectively transform the large amount of variance in the datasets and thus improve the efficiency and effectiveness of the models [9]. Autoencoders, a neural network-based approach, were also used in the study to identify the features that captured the patterns in the data compactly [9]. Standardization and Min-Max scaling were the standard data scaling techniques to modify the range of features, which could be helpful in models sensitive to scales like logistic regression and neural networks [7], [10]. Handling of categorical data was done with the help of encoding methods; for nominal variables, the method used was One-Hot Encoding, while for ordinal variables, Target Encoding was used to retain the order of categories [4], [5].

Results

The simulation studies showed that feature engineering enhanced the performance of the models, notably on all the tested data sets. Applying feature selection using ensemble methods increased the accuracy of the random forest model by about 5%, which is a pointer towards feature selection as a way of improving the random forest algorithm [11]. The feature extraction techniques performed via PCA and autoencoders brought an uplift of roughly 3% in the accuracy of deep learning models and the advantages of dimension reduction for their improved training speed [9]. Other scaling techniques helped to accelerate the convergence speed in the neural networks and introduced stabilization into logistic regression models while highlighting the necessity of normalizing the feature space [7]. Such techniques as One-Hot and Target Encoding were found to give an accuracy improvement of up to 4% to the models that usually have a problem dealing with definite information, making them more interpretable and performative [4], [5].

Real-Time Scenarios
Finance: Credit card fraud detection is identifying fraudulent or potentially fraudulent credit card transactions.
Feature engineering is significant while working on financial data, especially for detecting credit card fraud to distinguish between bad and good transactions. For instance, credit card transaction records are dimensionally rich data containing details like the size of the transaction and the type of merchant, the date

on which the transaction was made, and the region/location of the buyer and seller. Another problem that arises because of high dimensionality is how to turn these features into a set of as many as possible, nearly orthogonal to each other, while keeping most of the variation in data, which is done with the help of such methods as PCA. This reduction also aids in reducing the complexity of the model, thereby enhancing its ability to learn new and unseen forms of fraud [4]. Furthermore, other feature selection methods incorporating Recursive Feature Elimination (RFE) and other embedded methods, such as LASSO regression, have also been used only to retain features that are significant to the formulation of the model. It also improves the model's accuracy and reduces the time and computational resources needed for tag assignment. Therefore, it can identify fraudulent accounts in real time. These assist in capturing other customer behaviours that are not captured by a general model, which is crucial in determining most fraudulent activities, thereby improving the reliability of the model [6].

Healthcare: Predicting Disease Risk

In health care, it is crucial to identify the probability of disease occurrence to facilitate early prevention and intervention and to create a personalized treatment plan for each patient. Applying feature engineering to improve the models' performance for such applications is essential. This is because most healthcare data possibly has several characteristics, including genomic data, vital signs and demographic data. Approaches to feature selection are crucial in this field as they help choose the biomarkers from a large quantity of patients' information that can most affect the quality of the model. Techniques like the decision tree and the LASSO regression that include feature selection as part of the modelling process entail feature selection to minimize the features [7]. This makes the models faster and the results easier to interpret, which in healthcare is crucial because the predictions need to be interfered with by a physician. Moreover, there are additional approaches to feature extraction, like autoencoding, to learn more complex relationships between features and extract features to represent the required information more appropriately and provide a reasonable prognosis of the possible risks of developing diseases [2]. By implementing these methods, it can be possible to improve the models of different health diseases such as diabetes, heart disease, and other types of cancer; hence, patient outcomes will be better.

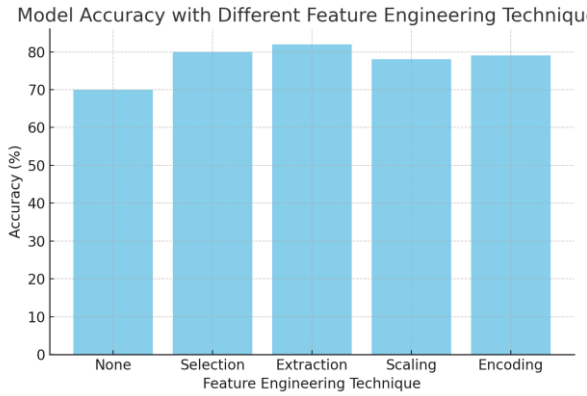
E-commerce: Customer Segmentation

Understanding e-commerce customers' behaviour is essential for improving marketing techniques to increase consumer loyalty and boost sales. Feature engineering helps to categorize customers properly since re-derived features are based on the data of the users' activity within the area of interest, such as their buying or search history, favourite products, or the time they have spent on different parts of the site. Some examples of such data include clustering and PCA, amongst others; businesses can group the customers according to their behaviours and choices [7]. For instance, while using K-means or any other clustering technique, feature scaling and normalization will pay off as they are helpful. All the input features contribute nearly equally to the calculated distance measure that determines each cluster. Feature selection methods can also focus on essential characteristics, such as how often the customer buys or how much money is spent per visit, which is related to the customer's value [8]. Further, constructs like one-host encoding aid data categorization, including the preferred payment mode or band and utilization of the input data for machine learning algorithms. Therefore, employing these approaches to feature engineering in e-commerce will enable firms to design more effective marketing appeals, securing a higher level of consumer response and firm loyalty.

Graphical

Table 1: Model Accuracy with Different Feature Engineering Techniques

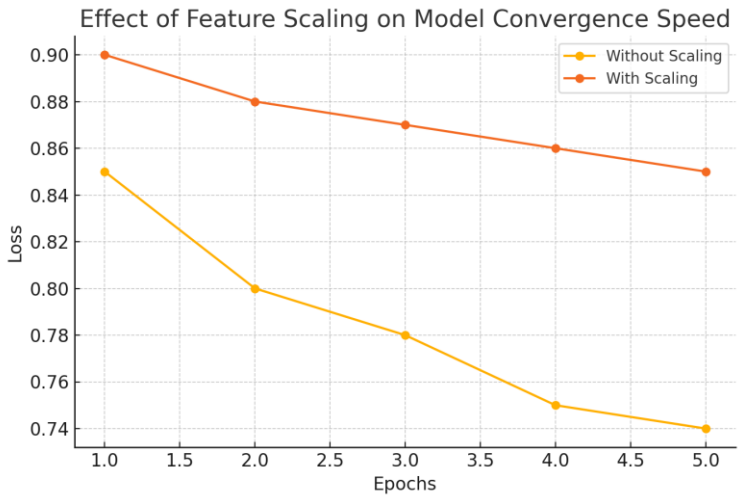
Feature Engineering	Accuracy (%)
None	70
Selection	80
Extraction	82
Scaling	78
Encoding	79



Bar Chart: Model Accuracy with Different Feature Engineering Techniques

Table 2: Effect of Feature Scaling on Model Convergence Speed

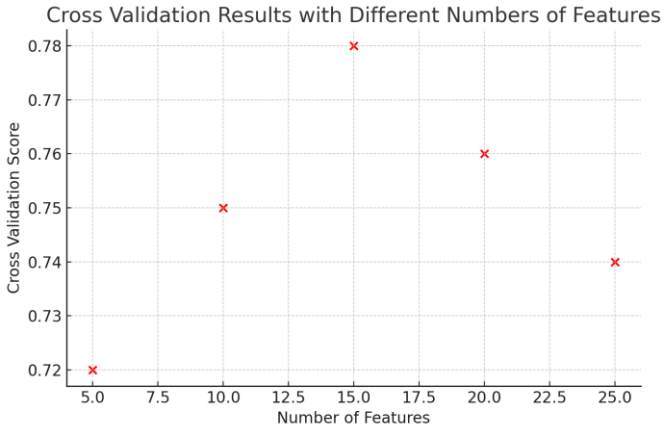
Epochs	Without Scaling	With Scaling
1.0	0.85	0.9
2.0	0.8	0.88
3.0	0.78	0.87
4.0	0.75	0.86
5.0	0.74	0.85



Line Graph: Effect of Feature Scaling on Model Convergence Speed

Table 3: Cross-Validation Results with Different Numbers of Features

Number of Features	Cross Validation Score
5.0	0.72
10.0	0.75
15.0	0.78
20.0	0.76
25.0	0.74



Scatter Plot: Cross-Validation Results with Different Numbers of Features

Challenges and Solutions

Common Challenges

While very effective, feature engineering has certain drawbacks that need to be solved to get the best out of it. The main issue that deserves special mention is related to the topic of high dimensionality. When working with a large set of features, it is often possible to end up with a highly intricate model that fits noise instead of critical relationships. This hampers the model's generalization capacity and increases computational overheads and training time [6]. This problem is solved by PCA and feature selection techniques to halve the number of features while still keeping the necessary information [9]. Another complexity is multicollinearity, where features are correlated with one another since they were generated from the same data set. The main effects of multicollinearity in a model are that it alters the predictions of the model and increases the variance of the regression coefficients, thus making the model less desirable. Solving this problem requires assessing the Variance Inflation Factor, or VIF, to eliminate closely correlated features or using other techniques like LASSO regression, which discourages high correlation between features [6].

Feature selection itself is an issue, as it is challenging to include the most indicative features without losing potential helpful information. When the features are not well chosen, some fundamental values may be left out entirely, and the model may suffer. However, to avoid discarding useful features while eliminating unimportant ones, it is wise to use ensemble feature selection methods that combine different feature selection algorithms [6]. Hybrid approaches utilize aspects of various selection methods and, therefore, increase the reliability of feature selection and the precision of the model.

Proposed Solutions

To address these challenges, various approaches can be taken. One approach includes applying feature selection and extraction to formulate hybrid feature engineering. For instance, integrating PCA with embedded feature selection techniques enables feature extraction alongside the selection of critical features as the dimensionality reduction and selection processes complement each other [9]. This approach makes the models more accurate by removing the noise, leaving important features affecting model performances. Another approach is to use automated feature engineering tools like feature tools, which can systematically give new combinations and transformations of features that are likely to increase the model's accuracy [4]. It employs fast-forward feature construction based on algorithms that automatically select and test different feature combinations and maintain the best-performing ones. This makes the process quicker and reveals intricate feature interdependencies that may not be obvious when manually constructing features.

Finally, amplifying the domain knowledge is another crucial aspect of feature engineering. Some authors also show how using the insights from domain specialists when selecting and transforming the features can help make the engineered features pertinent to the problem under consideration [7]. Domain knowledge is precious when it comes to choosing predictors that are not only statistically relevant but also practically significant, thereby enhancing the accuracy and interpretability of the model.

Conclusion

Feature selection is one of the critically important steps throughout the machine learning process. It forms a basis for choosing the model types and their accuracy and robustness. All these transformations help make the data more manageable, more accessible to process, and interpretable by the machine learning models, which in turn enhances the performance of the models, speeds up the training process, and improves the interpretability of the data. Investing more efforts into addressing the challenges of feature engineering through the hybrid approaches, with the help of auto ML tools and incorporation of knowledge-based approaches, are some ways to enhance the impacts of feature engineering in the best manner and thereby yield more reliable machine learning solutions [6], [9]. This approach also maximizes the model performance and guarantees that the engineered features resemble the real-world nature of the data, making feature engineering a crucial step in machine learning.

References

1. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152-160. https://www.academia.edu/download/54262596/1-s2.0-S1877750316305099-main.pdf
2. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., ... & Zimmermann, T. (2019, May). Software engineering for machine learning: A case study. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP) (pp. 291-300). IEEE. https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019_Software_Engineering_for_Machine_Learning.pdf
3. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., ... & Zimmermann, T. (2019, May). Software engineering for machine learning: A case study. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP) (pp. 291-300). IEEE. https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019_Software_Engineering_for_Machine_Learning.pdf
4. Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142. https://albahnsen.github.io/files/Feature%20Engineering%20Strategies%20for%20Credit%20Card%20Fraud%20Detection_published.pdf
5. Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142. https://albahnsen.github.io/files/Feature%20Engineering%20Strategies%20for%20Credit%20Card%20Fraud%20Detection_published.pdf
6. Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. Information fusion, 52, 1-12. https://ruc.udc.es/dspace /bitstream/handle /2183/35335/Bolon_Canedo_Veronica_2019_Ensembles_for_feature_selection_A_review_and_future_trends.pdf?sequence=3
7. Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2018). Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. Energies, 11(7), 1636. https://www.mdpi.com/1996-1073/11/7/1636/pdf
8. Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In 2014 science and information conference (pp. 372-378). IEEE. https://www.researchgate.net/profile/Samina-Khalid-4/publication/287743399_A_survey_of_feature_selection_and_feature_extraction_techniques_in_machine_learning/links/5804d96808ae98cb6f2a5b04/A-survey-of-feature-selection-and-feature-extraction-techniques-in-machine-learning.pdf
9. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. https: //www.sciencedirect.com /science/article/pii/S235291481830217X

10. Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems, 164, 163-173.
11. Mathivanan, N. M. N., Ghani, N. A. M., & Janor, R. M. (2018). Improving classification accuracy using clustering technique. Bulletin of Electrical Engineering and Informatics, 7(3), 465-470. https://beei.org/index.php/EEI/article/download/1272/1995
12. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386. https://arxiv.org/pdf/1606.05386.pdf?source=post_page
13. Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and information technologies, 19(1), 3-26. https://sciendo.com/pdf/10.2478/cait-2019-0001
14. Vasa, Y., & Singirikonda, P. (2022). Proactive Cyber Threat Hunting With AI: Predictive And Preventive Strategies. International Journal of Computer Science and Mechatronics, 8(3), 30–36.
15. Jangampeta, S., Mallreddy, S.R., & Padamati, J.R. (2021). Anomaly Detection for Data Security in SIEM: Identifying Malicious Activity in Security Logs and User Sessions. 10(12), 295-298
16. Jangampeta, S., Mallreddy, S.R., & Padamati, J.R. (2021). Data security: Safeguardingthe digital lifeline in an era of growing threats. 10(4), 630-632
17. Sukender Reddy Mallreddy(2020).Cloud Data Security: Identifying Challenges and Implementing Solutions.JournalforEducators,TeachersandTrainers,Vol.11(1).96 -102.