JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

The Role of Domain Knowledge in Feature Selection for Machine Learning

Abubakar Bello Bada*, A. B. Garko, Danlami Gabi, Musa S. Argungu

Department of Computer Science, Faculty of Physical Sciences, Kebbi State University of Science and Technology, Aliero.

***Corresponding Author**: abubakarbellobada@yahoo.com

ABSTRACT As a branch of artificial intelligence, machine learning focuses on using data and algorithms in order to make computers accomplish tasks without explicit programming. Feature selection is an essential phase in the machine learning process that has a huge impact on the performance of predictive models. Even though automated feature selection techniques are commonly used, incorporation of domain knowledge, an understanding and expertise in a specific area or field, in the process can bring about great improvement. This paper discusses the role domain knowledge plays in feature selection by showing how experts' perspectives help in identifying meaningful features that enhance model accuracy and applicability. It also comprehensively discusses ways of acquiring domain knowledge and incorporating it in feature selection. The paper also discusses the challenges that accompany incorporating domain knowledge in feature selection.

Keyword: Machine learning, Domain Knowledge, Deep Learning, Artificial Intelligence

I. INTRODUCTION

Machine Learning (ML) has changed several areas of human endeavor ranging from financial, education, healthcare, and others by providing predictive modelling and analytic tools. However, the success of these predictive modelling and analytic tools largely relies on the quality of features in the dataset used in training the models. Feature selection, which refers to identification, and selection of very important variables in a dataset is crucial in developing effective and efficient machine learning model (Melchior, 2022). Though there are many automated techniques for feature selection, incorporating domain knowledge into it can greatly improve the process (Groves, 2021). Domain experts can bring insights and expertise into selection of features that are not only significant statistically but also contextually meaningful (Himel & Sheng, 2022; Song et al., 2022). This paper explores the role of domain knowledge in feature selection by explaining how it aids selecting relevant features thereby improving performance of the model and ensures the results are interpretable. This paper emphasizes the importance of combining automated feature selection methods with domain knowledge by providing practical examples.

Feature selection is the process of selecting relevant features or a combination of features from a dataset. Selecting features or variables from the relevant dataset that can be used in training model is an important step in machine learning process. The main idea behind this feature selection is to find the features or variables that will be best in creating an efficient model (Aretove, 2022). The performance, efficiency, and interpretability of machine learning models is improved by feature selection technique. Using these techniques, features that are very important in dataset can be

VOLUME 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16615428

identified leading to reduction in overfitting and improvement in the generalization of the result. Applying feature selection techniques leads to development of models with better accuracy and are interpretable, this in turn leads to better decision making in areas of application (Habeeb, 2024). Feature selection process is essential for several reasons (Buyukkececi & Okur, 2023), this include:

(a). Improved Model Performance: Model's performance can be improved by removing or reducing overfitting. This can be achieved by removing data that is irrelevant and/or repetitive; this in turn enhances model's ability to generalize its result.

(b). Reduced Computational Cost: Feature selection leads to reduction in the number of features used in model training and fewer features means lower computational requirements.

(c). Enhanced Interpretability: Interpretability is very important in industries where decisions have major consequences e.g. finance, healthcare. Fewer set of features that are meaningful makes model easier to interpret. This interpretation is very significant in understanding the underlying data patterns and for gaining insights into how the model makes prediction, and understanding how the model makes its prediction helps ensure that the decisions made with those predictions are justifiable and trustworthy.

(d). Mitigation of Dimensionality: Data with high dimensionality usually leads to poor performance by the model. This dimensionality results when there is more features than observation in the data. Feature selection helps in mitigating the curse of dimensionality by reducing the number of features to only the needed ones.

II. RELATED WORKS

Automated feature selection techniques can be broadly classified into three categories (Menon, 2024): filter methods, wrapper methods, and embedded methods.

(a). Filter Methods: These methods use statistical techniques to determine how important a feature is irrespective of the model. These statistical methods include Pearson correlation, chi-square test, and mutual information.

(b). Wrapper Methods: These techniques evaluate feature subsets according on how well they work with a particular model. Examples of these methods include recursive feature elimination (RFE) and genetic algorithms.

(c). Embedded Methods: under these methods, features are selected during the model training. Regularization techniques like Lasso (L1) and Ridge (L2) regression are some of the examples of these methods.

Although these methods work quite well, they often fail to recognize the significance of features in relation to their industries. They rely solely on statistical computations, which can lead to the choice of some features that do not significantly contribute in the prediction process; this is why domain knowledge becomes invaluable in feature selection.

(a). Role of Domain Knowledge in Feature Selection

Domain knowledge refers to the expertise and insights that professionals possess about the specific context or field e.g., healthcare, finance, or real estate, from which the data is derived. It entails being familiar with the rules, details, and unique aspects of that field (GeeksforGeeks, 2021). Domain knowledge is very important in feature selection in machine learning. It serves as a map guiding through data jungle. It is essential for the accuracy of machine learning models, as it can help in avoiding common pitfalls and errors (Groves, 2022a). For instance, it can assist in avoiding data leakage by identifying and excluding features that are not available at the time of prediction.

VOLUME 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16615428

It can also aid in handling missing values by imputing new ones or dropping them (Melchior, 2022). Furthermore, the relevance of the chosen features can be influenced by incorporating domain knowledge into the feature selection process (Radovanović et al., 2019). Here are some ways in which domain knowledge contributes to feature selection:

(a). Identifying Relevant Features

In feature selection, domain knowledge plays a great role in identifying relevant features or variables. Through domain knowledge, domain experts can identify which features or variables can influence the outcome through their understanding of the underlying process (Guoyan et al., 2021). For example, suppose we want to predict housing prices using a Machine Learning model. An immediate issue that will be faced will be determining which features to use that are highly relevant and have high predictive power. This is where domain knowledge comes into play. Experts in the field, such as real estate agents, possess a deep understanding of factors that influence housing prices. This knowledge empowers them to remove features that have no much importance in determining the price e.g., color of the house, and identify features that are likely to influence the price of the house, such as proximity to amenities, and lower crime rate in the location. By incorporating these domain-informed features, the model can achieve a far more accurate prediction of housing prices compared to a model relying solely on generic features (Groves, 2022b). By focusing on these relevant features, domain knowledge helps in reducing the dimensionality of the dataset, making the model more efficient and potentially improving its accuracy. This targeted approach enhances the model's performance.

(b). Improving Feature Interpretability

Domain knowledge ensures that interpretation of features is easier because it makes sure that only relevant, meaningful, and understandable features that are within the context of the problem been solved get selected. This ensures the development of models that are more accurate, transparent, and easier to communicate and trust. Features selected with the help of domain knowledge are in most of the cases more interpretable (Dash et al., 2022). This is very important for fields where understanding the rationale behind model predictions is essential e.g. healthcare and finance. For instance, a financial analyst can explain why certain economic indicators are important for predicting market trends (Kumar et al., 2023).

(c). Enhancing Model Performance

By aiding feature selection process, domain knowledge significantly enhances model performance. It aides feature selection process in different ways e.g. by identifying the most relevant features and handling missing data and outliers. By identifying the most relevant features, it ensures that the model focuses on the most informative aspects of the data. This approach, which is a targeted one, helps in reducing dimensionality of the dataset and improves the efficiency and accuracy of the model by eliminating irrelevant or redundant features that could make the learning process complicated. In handling missing data and outliers, experts can help with better ways for imputation, transformation, or exclusion of such data points based on their knowledge of the domain (Somraj, 2021). This careful data processing through domain knowledge ensures the development of model on a good foundation, leading to a more reliable and robust performance.

Domain knowledge can be helpful in generating new features that capture important patterns in the data. For example, in natural language processing (NLP), linguistic experts can create features

VOLUME 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16615428

based on syntactic and semantic analysis, improving the performance of text classification models (CFI Team, 2024).

(d). Avoiding false Correlations

Applying domain knowledge in the selection of features helps in eliminating false relationship by making sure that the features selected have valid scientific relationship with the target variable. Selecting features using automated techniques can sometimes lead to selection of features that have low correlation with the target variable because of some chance occurrences. In this situation, domain knowledge plays a vital role as domain expertise provided by experts is used in distinguishing between coincidental correlation that happens due to some random factors and genuine correlation. This helps in preventing the inclusion of features that appear statistically significant but do not have any meaningful connection to the problem being solved (Warepam, 2023).

Through focusing on features that are backed by established theories and empirical evidence, the risk of incorporating misleading variables that could lead to incorrect predictions by the model is reduced (Explorium, 2023). This selection process enhances the model's reliability and validity.

III. METHODOLOGIES FOR INCORPORATING DOMAIN KNOWLEDGE

There are many techniques for integrating domain knowledge into feature selection. They include the following:

(a). Manual Feature Engineering

This is a process of feature creation and selection through domain expertise only. This process may encompass conversion of raw data into meaningful features, combination of individual variables to form compound features, and the application of domain specific knowledge to aid feature selection (Melchior, 2022). For instance, in a study conducted to predict readmission of patients into hospital, domain experts recognized certain characteristics features of patients' e.g. previous admission history, comorbidities, and medication compliance as important features. Automated technique alone may not necessarily pick these features even though they have significant impact on the accuracy of the model in this instance (Davis et al., 2022).

(b). Hybrid Feature Selection Methods

The combination of automated techniques with domain knowledge in feature selection to utilize the strength of both is called Hybrid feature selection method. There are two approaches to implementing this method; one is that automated methods are used to generate initial features that are then reviewed and refined by domain experts. Another approach is that domain knowledge is used to define constraints and guide the search process of automated method (Dash et al., 2022; Melchior, 2022). Domain experts validate the importance of features identified by automated methods ensuring that they are relevant to the problem at hand. They can also help in creating new features that accurately represent the domain and ignore the ones that even though are statistically significant, have no practical value. This feature selection approach ensures that the derived features are meaningful leading to the development of models that are accurate, stable, and easy to interpret (Melchior, 2022).

In an example of a study that involved financial fraud detection, initial features were obtained with machine learning algorithms, and then the features were refined by domain experts based on their understanding of the patterns of frauds. This led to fewer false positive classifications and an overall better model stability (Park et al., 2022).

VOLUME 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16615428

(c). Domain-Specific Ontologies and Taxonomies

Domain-specific ontologies and taxonomies provide systematized approach for organizing knowledge in a specified area. These systematized approaches hierarchically classify concepts and relationships to enhance systematic incorporation of domain knowledge in feature selection process. By using ontologies and taxonomies, feature selection can be done using already established hierarchies of domain-specific terms and their interrelations, ensuring that relevant features that aligned with established domain principles are selected. This approach aids in finding key features that are in tandem with relevant concepts within the domain thereby eliminating grounds for including irrelevant or confusing variables (Melchior, 2022). It is particularly relevant in areas such as medicine and biology where relationship between entities, though difficult, has to be represented.

For instance, in genomic data analysis, it is common to use ontologies like the Gene Ontology (GO) to create features that represent biological processes, cellular components, and molecular functions. These features have been found to improve the performance of models in classifying gene-associated diseases (Pastor et al., 2021).

IV. METHODOLOGIES OF ACQUIRING DOMAIN KNOWLEDGE

Acquiring domain knowledge is essential for effectively applying ML techniques to real-world problems. Here are some methodologies to acquire domain knowledge in ML (Mende et al., 2023; Xie et al., 2021):

a) Review of Literature and Research: reading of research papers and review articles that focus on the area of interest will give an insight into methodologies used in the domain. Study of domain-specific books that cover concept, terminologies, and problems specific to the field and complimenting it with ML-specific literature to understand how ML techniques are applied to these problems will help in acquiring the required knowledge.

b) Collaboration with Domain experts: collaborating with experts e.g. doctors, engineers, in domain of interest and working closely with them can provide a good view into practical challenges associated with the domain that cannot be seen from the data. Formal and informal interview with the expert is another way of collaboration and it is crucial to treat these experts with respect and show them that you understand their field of expertise.

c) Enroll in Domain-specific courses and certifications: enrolling in educational classes related to the domain of interest is a strategy to obtain domain knowledge, most especially courses that combine machine learning with domain-specific knowledge e.g., AI in healthcare.

d) Networking and Community engagement: Engaging with professionals, developing a robust network both online and offline that focus on the application of ML in specific domains is another way of acquiring domain-specific knowledge.

e) Iterative learning: ML and domain-specific knowledge are continually evolving. Stay updated by regularly reviewing new research, tools, and methodologies. Use online platforms like Coursera etc. to refresh your knowledge. Also, apply ML techniques to domain-specific problems and seek feedback from domain experts. The feedback from the experts will deepen the understanding of the domain.

V. CHALLENGES OF INCORPORATING DOMAIN KNOWLEDGE IN FEATURE SELECTION

Even though incorporating domain knowledge into feature selection in machine learning has many advantages, it also comes with its own challenges. Potential for bias and overfitting is one of those

VOLUME 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16615428

challenges. Even with their expertise, experts in the field may give more emphasis on some features that they believe are important while downplaying others however significant they may be, hence narrowing the focus. This can result in a biased prediction by the model thereby reducing its generalizability (Himel & Sheng, 2022). Furthermore, depending too much on domain knowledge may lead to development of models that do well on training data but poorly on new data, as the selected features do not capture the underlying patterns fully. Finding a middle ground between domain knowledge and data-driven insight is very important in handling these risks (Barbierato & Gatti, 2024).

Lack of communication between domain experts and data scientists is another challenge faced in incorporating domain knowledge in feature selection (Wuest et al., 2016). In order to select features effectively, it is necessary for all the parties (domain experts and data scientists) involved to collaborate well and have mutual understanding among themselves. This is very difficult to achieve due to different perspective, terminologies, and methodologies. Domain experts may find it difficult to fashion their knowledge in a way that it will be easily converted into machine learning features, at the same time, it will be challenging for data scientists to interpret and incorporate the expertise into models (Himel & Sheng, 2022; Xie et al., 2021). Removing this barrier needs dialogue and development of interdisciplinary skills.

VI. CONCLUSION

Domain knowledge is essential in feature selection for developing effective and interpretable machine learning models. By using professional expertise in a specific area, relevant and meaningful features can be elicited which will improve interpretability and performance of model. Even though there are challenges in integrating domain knowledge in feature selection, achieving robust and reliable predictive model as we look into the future lies in seamless combination of domain knowledge and automated feature selection techniques. In conclusion, incorporating domain knowledge in feature selection process for machine learning is at the same time a challenge and an asset. This paper has highlighted the importance of expertise in a particular domain in improving the accuracy and interpretability of models. Thus, when such expertise is used in feature selection, features with most predictive tendency can be identified and given high priority thereby enhancing the performance of the model. This process nonetheless has its own challenges such as biases, which may arise during data mining processes, overfitting, and gap in communication that may occur between data scientists and domain experts. In dealing with these challenges, there should always be equilibrium between incorporating data driven techniques along with domain knowledge for creating collaborative environments where continuous learning takes place.

In the future, incorporating domain knowledge in feature selection lies on improved collaboration between disciplines and the use of advanced technologies. Automated feature engineering tools, explainable AI (XAI), and transfer learning techniques are set to close down the gap between expertise in domain areas and machine learning. The improvements will facilitate faster and smoother process for selecting features. Thus, due to evolving nature of the field, it is important to have an approach that continuously validate the features to make sure they are still relevant due to arrival of new data and ever-changing domain knowledge. In the end, properly incorporating domain expertise in selection of features would enable us not only to achieve models that are more accurate, interpretable but also promote practical application as well as innovation across various fields.

VOLUME 1, Issue 3, 2025
https://doi.org/10.5281/zenodo.16615428

REFERENCES

[1] Aretove. (2022). *Importance of Feature Selection in Machine Learning*. https://www.aretove.com/importance-of-feature-selection-in-machine-learning

[2] Barbierato, E., & Gatti, A. (2024). The Challenges of Machine Learning: A Critical Review. *Electronics (Switzerland)*, *13*(2). https://doi.org/10.3390/ELECTRONICS13020416

[3] CFI Team. (2024). *Domain Knowledge (Data Science) - Overview, Case Study*. https://corporatefinanceinstitute.com/resources/data-science/domain-knowledge-data-science/

[4] Dash, T., Chitlangia, S., Ahuja, A., & Srinivasan, A. (2022). A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports 2022 12:1*, *12*(1), 1–15. https://doi.org/10.1038/s41598-021-04590-0

[5] Davis, S., Zhang, J., Lee, I., Rezaei, M., Greiner, R., McAlister, F. A., & Padwal, R. (2022). Effective hospital readmission prediction models using machine-learned features. *BMC Health Services Research*, *22*(1), 1–10. https://doi.org/10.1186/S12913-022-08748-Y/FIGURES/3

[6] Explorium. (2023). *Domain Knowledge in Data Science: Data Science for Business*. https://www.explorium.ai/blog/machine-learning/domain-knowledge-in-data-science-are-your-models-ready-for-business/

[7] GeeksforGeeks. (2021). *Role of Domain Knowledge in Data Science*. https://www.geeksforgeeks.org/role-of-domain-knowledge-in-data-science/

[8] Groves, W. (2021). *Using Domain Knowledge to Systematically Guide Feature Selection*.

[9] Groves, W. (2022a). *Using Domain Knowledge to Systematically Guide Feature Selection*.

[10] Groves, W. (2022b). *Using Domain Knowledge to Systematically Guide Feature Selection*.

[11] Guoyan, L., Yuan, C., Kamarthi, S., Moghaddam, M., & Jin, X. (2021). Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis. *Journal of Manufacturing Systems*, *60*, 692–706. https://doi.org/10.1016/J.JMSY.2021.07.007

[12] Habeeb, M. J. (2024). *Feature Selection Techniques in Machine Learning*. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

[13] Himel, D. G., & Sheng, V. S. (2022). *A Roadmap to Domain Knowledge Integration in Machine Learning*.

[14] Kumar, A., Hooda, S., Gill, R., Ahlawat, D., Srivastva, D., & Kumar, R. (2023). Stock Price Prediction Using Machine Learning. *Proceedings of International Conference on Computational Intelligence and Sustainable Engineering Solution, CISES 2023*, 926–932. https://doi.org/10.1109/CISES58720.2023.10183515

[15] Melchior, M. (2022). *Incorporating Domain Knowledge for Learning Interpretable Features*. https://doi.org/10.5445/IR/1000150142

[16] Mende, H., Frye, M., Vogel, P. A., Kiroriwal, S., Schmitt, R. H., & Bergs, T. (2023). On the importance of domain expertise in feature engineering for predictive product quality in production. *Procedia CIRP*, *118*, 1096–1101. https://doi.org/10.1016/J.PROCIR.2023.06.188

[17] Menon, K. (2024). *Feature Selection Techniques in Machine Learning*. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

https://doi.org/10.5281/zenodo.16615428

[18] Park, S., Kim, S., & Cha, M. (2022). *Knowledge Sharing via Domain Adaptation in Customs Fraud Detection*. www.aaai.org

[19] Pastor, Ó., León, A. P., Reyes, J. F. R., Garciá, A. S., & Casamayor, J. C. R. (2021). Using conceptual modeling to improve genome data management. *Briefings in Bioinformatics*, *22*(1), 45–54. https://doi.org/10.1093/BIB/BBAA100

[20] Radovanović, S., Delibašić, B., Jovanović, M., Vukićević, M., & Suknović, M. (2019). A Framework for Integrating Domain Knowledge in Logistic Regression with Application to Hospital Readmission Prediction. *Https://Doi.Org/10.1142/S0218213019600066*, *28*(6). https://doi.org/10.1142/S0218213019600066

[21] Somraj, S. (2021). *Domain Knowledge - Perhaps the Most Important Skill in Data Science*. https://towardsdatascience.com/the-second-most-important-skill-to-have-as-a-data-scientist-2311f9efa143

[22] Song, W., Hou, Z., Gu, W., Afgan, M. S., Cui, J., Wang, H., Wang, Y., & Wang, Z. (2022). Incorporating domain knowledge into machine learning for laser-induced breakdown spectroscopy quantification. *Spectrochimica Acta - Part B Atomic Spectroscopy*, *195*. https://doi.org/10.1016/J.SAB.2022.106490

[23] Warepam, R. (2023). *Why "Domain Knowledge" is IMPORTANT for Data Science Jobs?* https://medium.com/dare-to-be-better/why-domain-knowledge-is-important-for-data-science-jobs-8195a5d3e84e

[24] Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production and Manufacturing Research*, *4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

[25] Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., & Yu, S. (2021). A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, *69*, 101985. https://doi.org/10.1016/J.MEDIA.2021.101985

https://doi.org/10.5281/zenodo.16615428