Feature Selection in the Age of Large Language Models: Insights from DeepSeek

Marija Đukić
University of Belgrade, Faculty of Organizational Sciences
Belgrade, Serbia *marija.djukic@fon.bg.ac.rs*

Aleksandra Sretenović
University of Belgrade, Faculty of Organizational Sciences
Belgrade, Serbia *aleksandra.sretenovic@fon.bg.ac.rs*

Abstract

Feature selection has great importance for simplifying machine learning and improving computational efficiency, especially when working with high-dimensional datasets. The rise of Large Language Models (LLMs) offers new opportunities in selecting predictive features. This paper aims to evaluate the potential of LLMs for feature selection tasks and examine whether a hybrid approach can lead to improved predictive performance. Using the DeepSeek-R1 model on publicly available datasets, the results show that LLM-driven feature selection holds significant promise. Furthermore, the performance of hybrid approaches highlights the value of LLMs as a complementary tool to traditional feature selection methods. Across the experiments, the hybrid approach either achieved the highest performance or ranked among the top-performing methods.

Keywords: Large Language Models, Feature Selection, Classification, DeepSeek

1. Introduction

With large and complex datasets becoming widely available, the potential to develop high-performing machine learning models increases. These datasets can uncover patterns that were previously difficult to detect because of their high dimensionality, heterogeneity, and complex interdependence [2]. However, this complexity comes with a challenge: selecting the most relevant elements to build effective models.

Feature selection is a process of identifying a subset of the most informative features from a larger set of features. The goal is to build simpler models, easier to understand and improve computational efficiency, particularly in high-dimensional scenarios [8]. Traditional feature selection methods are classified into three categories: filter, wrapper, and embedded. Filter methods rank features by assigning a score to each feature independently of a learning model. Typically, either the top N features with the highest scores are selected, or all features that exceed a predefined threshold. Wrapper methods evaluate subsets of features by fitting a supervised learning model to each subset and assessing performance using a predefined metric [10]. These methods rely on statistical metrics that may not capture complex, nonlinear relationships within the data. Moreover, they often require significant computational resources and training data [3]. While existing selection methods show good performance in data-rich contexts, there is a growing need for efficient feature selection with limited, or no training samples. This need is particularly noticeable in fields with sensitive data such as healthcare, where privacy concerns restrict data sharing, thus complicating the feature selection process [13].

The advent of Large Language Models (LLMs) introduced new opportunities for various machine learning tasks. Leveraging vast training data and well-designed prompting techniques, LLMs have demonstrated notable zero-shot and few-shot performance in tasks such as information extraction [4], code generation [14], and data analytics [1]. Different from traditional data-driven approaches, LLMs offer a novel perspective for feature selection through their semantic reasoning capabilities and in-context learning potential.

The objective of this research is to evaluate the capabilities of large language models

in feature selection in predictive modeling. For the case study, the DeepSeek-R1 model is used with publicly available employee attrition datasets. This paper seeks to address the following research questions:

RQ1: How effectively can LLMs identify relevant features for prediction tasks?

RQ2: Can a hybrid approach that combines LLM capabilities with traditional feature selection methods improve model predictive performance?

This paper is organized as follows: Section 2 provides a review of related work in the field. An overview of the DeepSeek-R1 model is given in Section 3, and description of the case study in Section 4. Section 5 offers discussion of the results, while Section 6 concludes the paper with directions for future research.

2. Related Work

Due to their ability to understand and generate text across diverse topics, large language models have found applications in numerous tasks. Their potential for feature selection has been the subject of recent research, with encouraging findings.

Authors of [9] distinguish between two categories of LLM-based feature selection methods: text-based, that utilizes descriptive contextual information, and data-driven, which require access to numerical sample values. The results show that text-based feature selection with LLMs is more consistent and stable in low-resource settings. The use of LLMs for transforming tabular data to improve the performance of machine learning models in binary classification tasks is explored in [5]. The combination of LLM-driven feature selection with data transformation significantly improved classification accuracy. Although peer-reviewed research in this area remains limited, recent experimental studies have offered promising results. Authors of [7] showed that LLMs can recognize the most predictive features even when they are given only the names of the input features and a description of the prediction task. In [11], a hybrid approach called LLM4FS is introduced that involves utilizing traditional methods performed by LLM to select relevant features.

This research uses both data-driven and text-based approaches for LLM-driven feature selection expanding on [9]. The hybrid approach combines the outputs of the LLM with traditional feature selection techniques, unlike [11] where the LLM executed traditional methods. The aim is to leverage the semantic capabilities of LLMs and the statistical robustness of traditional methods.

3. DeepSeek LLM

DeepSeek-R1 is an open-source large language model developed by DeepSeek AI that introduces novel approaches in its architecture and training process. It employs a four-phase training approach centered on rule-based reinforcement learning, in contrast to other LLMs that usually adhere to a three-stage training pipeline centered on supervised fine-tuning [6]. Another feature of DeepSeek-R1 is its use of Knowledge Distillation (KD), machine learning technique that transfers knowledge from a large, complex model to a smaller, more efficient one [15]. That way, KD improves inference speed and reduces computational costs without compromising the performance of the larger model. In this study, DeepSeek is selected due to its open-source nature, cost-effectiveness, and strong reasoning capabilities.

4. Case Study

The potential of LLMs in identifying relevant features is evaluated on two publicly available classification datasets from Kaggle. The first dataset comprises 24 features[1], and the second 35[2], allowing for analysis across different dimensionalities. The first dataset was sampled to a comparable size to the second, ensuring that differences in dimensionality could be analyzed without the impact of dataset size. Three feature selection strategies were used for each dataset: (1) traditional methods, (2) LLM-driven approach, and (3)

[1] https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors

[2] https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset

hybrid method integrating both. The Altair RapidMiner was used for the implementation, as it provides a modular, visual interface for building and executing workflows.

LLM-driven feature selection was performed with DeepSeek-R1 model, which was used to identify relevant features through both text-based and data-driven approaches. For the text-based approach, two prompts were used: feature ranking and feature importance. The importance prompt was tested under both zero-shot (no examples provided) and few-shot (limited examples provided) learning conditions. Hybrid approach guided by [12] was tested, with the aim of using the advantages of both approaches. The DeepSeek-R1 model was prompted to perform an initial semantic analysis of the features, placing them into groups based on their semantic similarity, thereby creating a hierarchical structure. Traditional feature selection methods, namely Gini Index and Relief, were used to evaluate and score individual features within each derived category. Only the feature with the highest score from each category was selected for the classification model.

To evaluate the predictive performance of the selected features, Random Forest and Logistic Regression classification algorithms were used. The classification model's performance is evaluated using the Area Under the ROC Curve (AUC) and the F1-score. AUC was used to assess the overall performance of the models in distinguishing between the positive and negative classes. The F1-score provides insight into class-specific performance, especially if the positive class is underrepresented.

5. Results and Discussion

The experimental results reveal performance differences for various feature selection approaches (Table 1). For the first dataset (24 features), traditional methods performed highly across both classifiers, outperforming LLM-driven approaches. The hybrid approach achieved the highest scores, with the combination of LLM and Gini Index reaching an AUC of 0.788 for Random Forest and 0.833 for Logistic Regression. These AUC values demonstrate the models' discriminative ability to distinguish between positive and negative attrition cases, indicating that the hybrid approach can effectively identify features which contribute to better prediction performance. In terms of F1-score, the LLM+Gini Index combination achieved 69.44 for Random Forest, closely approaching the highest score obtained with Relief (69.68), while LLM+Relief achieved the highest F1-score overall (74.96) for Logistic Regression. These findings support the notion that integrating the semantic understanding of LLMs with the statistical rigor of traditional methods can lead to improved predictive performance. In addition to improving predictive accuracy, feature selection helps develop practical and efficient machine learning models. Reducing the number of features simplifies the model and enhances interpretability. Additionally, simpler models require less training and inference time, making them better suited for deployment in resource-constrained settings or real-world applications. In this study, the hybrid approach retained strong performance while utilizing fewer features, showing potential in balancing accuracy, interpretability, and computational efficiency.

Table 1. Results of dataset with 24 features

Feature selection method \ Classification model		Random Forst		Logistic Regression	
		AUC	F1	AUC	F1
No Feature Selection		0.780	67.15	0.831	74.11
Traditional (Filter)	Gini Index (5 features)	0.771	68.98	0.807	72.14
	Relief (5 features)	0.703	64.67	0.755	70.77
	Gini Index (10 features)	0.784	68.90	0.819	73.71
	Relief (10 features)	0.785	**69.68**	0.795	71.67
Traditional (Wrapper)	Forward Selection	0.757	65.20	0.764	70.46
LLM-driven	Ranking (5 features)	0.560	45.51	0.621	64.15
	Ranking (10 features)	0.565	45.33	0.624	62.47
	Importance (zero-shot)	0.623	57.72	0.678	66.86
	Importance (few-shot)	0.646	61.08	0.664	65.73
	Data-driven (5 features)	0.580	44.82	0.614	63.60
	Data-driven (10 features)	0.667	56.84	0.705	64.97
Hybrid approach	LLM + Gini Index (5 features)	**0.788**	69.44	**0.833**	72.50
	LLM + Relief (5 features)	0.768	68.65	0.802	**74.96**

The experimental results (Table 2) for the second dataset (35 features) reflect that LLM-driven approaches show competitive capabilities, while traditional filter methods maintain strong predictive performance. Few-shot importance prompt with Logistic Regression achieved the highest AUC of 0.818. The hybrid approach delivered robust performance; the LLM+Gini combination achieved the highest overall AUC for Random Forest, reaching 0.815. In this case, the hybrid approach utilized 10 features, selected as the top-ranked across 7 derived feature categories. Regarding F1-scores, traditional methods achieved the highest performance using 15 features for both classifiers.

Table 2. Results of dataset with 35 features

Feature selection method \ Classification model		Random Forest		Logistic Regression	
		AUC	F1	AUC	F1
No Feature Selection		0.791	54.37	0.823	65.03
Traditional (Filter)	Gini Index (10 features)	0.752	62.36	0.761	58.28
	Relief (10 features)	0.757	60.50	0.776	60.34
	Gini Index (15 features)	0.787	64.39	0.795	**65.78**
	Relief (15 features)	0.778	**66.44**	0.777	63.57
Traditional (Wrapper)	Forward Selection	0.738	57.87	0.754	57.64
LLM-driven	Ranking (10 features)	0.780	64.45	0.795	56.54
	Ranking (15 features)	0.793	55.32	0.808	64.55
	Importance (zero-shot)	0.772	54.01	0.783	60.29
	Importance (few-shot)	0.810	63.27	**0.818**	62.70
	Data-driven (10 features)	0.772	61.05	0.793	55.68
	Data-driven (15 features)	0.792	60.35	0.793	56.67
Hybrid approach	LLM + Gini Index (10 features)	**0.815**	63.66	0.805	62.53
	LLM + Relief (10 features)	0.801	64.30	0.809	64.85

Compared to the first dataset, the results suggest that LLM-based feature selection scales more effectively in larger feature space. This highlights the potential value of LLMs for high-dimensional feature selection tasks, where traditional methods may encounter scalability challenges. Complex real-world datasets, like those in the medical or financial domains, often have significantly high dimensionality, sometimes involving hundreds or thousands of features. Therefore, LLMs may offer clear benefits in such extreme-scale scenarios; however, empirical validation on these high-dimensional datasets remains an important direction for future research. Although this case study offers valuable insights into LLM-driven feature selection, several limitations should be acknowledged. First, the analysis is based on a small datasets, which may affect performance. Second, the prompting strategy has a significant impact on how effective the LLM-driven approach is, and different prompt designs may produce different results.

6. Conclusion

The advent of LLMs introduced new opportunities for various machine learning tasks. This research explored the potential of LLMs for feature selection tasks.

In response to the first research question, the LLM-driven approach illustrated promising potential in feature selection, although its performance partially matches that of traditional methods. The current LLM-driven approach may require more sophisticated prompting or domain-specific fine-tuning. However, it showed improved performance on dataset with more features, demonstrating its potential for high-dimensional datasets.

In response to the second research question, the findings show that hybrid approach can result in improved predictive performance. For the dataset with fewer features, the hybrid method achieved the highest overall performance and ranked among the top-performing methods for the second dataset. Importantly, the hybrid approach achieved comparable or superior performance using fewer features, thereby simplifying the model and improving interpretability.

This research demonstrates the potential of LLM-driven feature selection and hybrid approaches emphasizes the value of LLMs as a complementary tool to traditional feature selection methods. Future research should explore the scalability of LLM-driven feature

selection and test this approach to more complex, real-world datasets with hundreds of features. To improve LLM performance, more sophisticated prompting strategies can be designed, and techniques for finding a balance between semantic abilities of LLMs and statistical rigor of traditional methods.

References

1. Almheiri, S. M. A. A., AlAnsari, M., AlHashmi, J., Abdalmajeed, N., Jalil, M., Ertek, G.: Data Analytics with Large Language Models (LLM): A Novel Prompting Framework. In: International Conference on Business Analytics in Practice, pp. 243-255. Springer Nature Switzerland, Cham (2024, January)
2. Bastarache, L., Brown, J. S., Cimino, J. J., Dorr, D. A., Embi, P. J., Payne, P. R., ... Weiner, M. G.: Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. Learn. Health Syst. 6(1), e10293 (2022)
3. Dhal, P., Azad, C.: A comprehensive survey on feature selection in the various fields of machine learning. Appl. Intell. 52, 4543–4581 (2021). https://doi.org/10.1007/s10489-021-02550-9
4. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., ... Feder, A.: LLMs accelerate annotation for medical information extraction. In: Machine Learning for Health (ML4H), pp. 82-100. PMLR, December (2023)
5. Haque, R., Goh, H. N., Ting, C. Y., Quek, A., Hasan, M. R.: Leveraging LLMs for optimised feature selection and embedding in structured data: A case study on graduate employment classification. Comput. Educ.: Artif. Intell. 8, 100356 (2025)
6. Hayder, W. A.: Highlighting DeepSeek-R1: Architecture, Features and Future Implications. Int. J. Comput. Sci. Mob. Comput. (IJCSMC) 14(2), 1-13 (2025). https://doi.org/10.47760/ijcsmc.2025.v14i02.001
7. Jeong, D. P., Lipton, Z. C., Ravikumar, P.: LLM-Select: Feature Selection with Large Language Models. arXiv e-prints, arXiv-2407 (2024)
8. Kaur, A., Guleria, K., Trivedi, N. K.: Feature selection in machine learning: Methods and comparison. In: 2021 Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), pp. 789-795. IEEE, March (2021)
9. Li, D., Tan, Z., Liu, H.: Exploring large language models for feature selection: A data-centric perspective. ACM SIGKDD Explor. Newsl. 26(2), 44-53 (2025)
10. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Comput. Surv. (CSUR) 50(6), 1-45 (2017)
11. Li, J., Xiu, X.: LLM4FS: Leveraging Large Language Models for Feature Selection and How to Improve It. arXiv preprint arXiv:2503.24157 (2025)
12. Radovanović, S., Delibašić, B., Vukanović, S.: Combining LLM and DIDEX method to predict Internal Migrations in Serbia. In: Human-Centric Decision and Negotiation Support for Societal Transitions (2024)
13. Remeseiro, B., Bolon-Canedo, V.: A review of feature selection methods in medical applications. Comput. Biol. Med. 112, 103375 (2019)
14. Wang, J., Chen, Y.: A review on code generation with LLMs: Application and evaluation. In: 2023 IEEE Int. Conf. Med. Artif. Intell. (MedAI), pp. 284-289. IEEE, November (2023)
15. Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., ... Chen, Y.: Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application. ACM Trans. Intell. Syst. Technol., 1–27 (2024). https://doi.org/10.1145/3699518

Appendix

The prompts used in the study are available at: github.com/m-djukic/LLM-FS-prompts