Journal of Economic Theory and Business Management
Journal Home: https://jetbm.suaspress.org/ | CODEN: JETBAU
Vol. 1, No. 6, 2024 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

Machine learning and Feature Selection: Applications in Business Management

CHEN, Siyao [1*] DAIYA, Abhit [1] DEB, Rubhrat [2]

[1] University of Connecticut, USA
[2] Wildlife Institute of India, India

**** CHEN, Siyao is the corresponding author, E-mail: suiyaoch@gmail.com***

Abstract: In recent years, we have a higher demand for machine learning models in the field of economics and business managment. We have a higher demand for quality of the features used for training. In this process, feature selection plays a key role in identifying the most meaningful features from a dataset while we perfrom various business tasks. Feature selection is not just a technical exercise; it also has profound implications for the transparency and explainability of machine learning models. This study aims to be a valuable resource for both academic and industry experts, offering insights that connect theoretical knowledge with practical implementation. And this paper also highlights the potential applications and significance of feature selection across industries like business, finance, and other real-world scenarios. And it aims to explore deep in the feature selection, showing that its impact on model performance and its role in various domains.

Keywords: Machine Learning, Feature Selection, Business Management.

Disciplines: Business. **Subjects:** Business Strategy.

DOI: https://doi.org/10.70393/6a6574626d.323438 **ARK:** https://n2t.net/ark:/40704/JETBM.v1n6a05

1 INTRODUCTION

In the era of big data, machine learning models are increasingly being used to mine patterns and insights from massive datasets. However, the effectiveness of these models often depends on the quality of the features used for training. Feature selection is a key preprocessing step that involves identifying the most meaningful features from a dataset while eliminating redundant or irrelevant ones. This process not only enhances the model's interpretability but also improves its computational efficiency and predictive performance.

High-dimensional data is common in fields such as genomics, image processing, and natural language processing, presenting unique challenges to machine learning algorithms. An excess of irrelevant features can lead to overfitting, where the model performs well on the training data but poorly on unseen data. Feature selection addresses this challenge by reducing dimensionality, thereby aiding in the creation of robust and generalizable models. Additionally, it helps to decrease computational costs and the time required for model training, making it an indispensable tool for researchers and practitioners.

The objective of this study is to conduct a thorough examination of feature selection methods within the field of machine learning, emphasizing their theoretical foundations, procedural frameworks, and practical applications. The paper systematically explores a range of techniques, including filter, wrapper, and embedded approaches, to clarify their comparative advantages, inherent limitations, and their suitability for various data scenarios.

The scope of this research also extends to assessing the latest developments and emerging trends in feature selection, particularly its synergy with deep learning and the field of explainable AI. Furthermore, the paper highlights the multifaceted application of feature selection across industries such as healthcare, finance, and image recognition systems, supported by case studies from real-world scenarios. Most importantly, this study aims to be a valuable resource for both academic and industry experts, offering insights that connect theoretical knowledge with practical implementation.

In the realm of artificial intelligence, feature selection stands as a pivotal component in the construction of effective machine learning models. It is the process of selecting a subset of relevant features for use in model construction, which is crucial for managing the information overload presented by big data. By carefully selecting features, we can enhance the model's ability to generalize from the training data to new, unseen data. This step is essential in preventing overfitting and ensuring that the model remains accurate and reliable in diverse situations. The study at hand aims to delve into the intricacies of feature selection, evaluating its impact on model performance and its role in various domains.

The importance of feature selection is further underscored by its ability to streamline the machine learning

process. In high-dimensional data scenarios, where the number of features may exceed the number of observations, feature selection becomes a necessity. It helps in identifying the most informative features that contribute significantly to the predictive power of the model. This not only leads to more efficient models but also aids in uncovering the underlying structure of the data. The study will scrutinize how feature selection techniques can be tailored to specific contexts, such as different industries and types of data, to maximize their effectiveness.

Feature selection is not just a technical exercise; it also has profound implications for the transparency and explainability of machine learning models. It has proven its ability to enhance model interpretability, which can greatly improve system explainability[1]. The researchers showed that by reducing the number of features and prioritizing those considered more relevant from a domain expert's perspective, feature selection techniques contribute to the development of more explainable artificial intelligence (XAI) models. As the field moves towards more accountable AI, the ability to understand and justify model decisions becomes increasingly important. Feature selection plays a critical role in this regard by reducing complexity and focusing on the most relevant aspects of the data. The comprehensive review presented in this study will explore the latest methodologies in feature selection and their implications for creating more transparent and explainable AI systems[2]. By doing so, it seeks to equip researchers and practitioners with the knowledge to build models that are not only performant but also trustworthy.

2 LITERATURE REVIEW

Feature selection is a core component in data analysis, involving the process of filtering out the most critical features for model prediction from complex datasets[3]. In machine learning projects, this step is crucial for optimizing model performance, especially when dealing with high-dimensional datasets that contain thousands of features. By eliminating irrelevant features, we can reduce the complexity of the model, improve its operational efficiency, and enhance its interpretability. In machine learning, the goal of feature selection is to distill the subset of features that best represent the essence of the data. Feature selection is not just a technical exercise; it also has profound implications for the transparency and explainability of machine learning models.This not only reduces the computational resources required for model training but also enhances the model's performance on new data. Feature selection helps models avoid overfitting by discarding features that have little impact on the prediction outcome, ensuring that the model maintains stable and accurate predictive capabilities in practical applications. The importance of feature selection is evident across various domains, especially in those with numerous data features, such as genomics, financial analysis, and image processing. In these fields, datasets may contain thousands of features, but not all contribute significantly to predicting the target outcome. Feature selection techniques can help us sift through these vast amounts of features to identify the truly important information, thereby constructing more precise and efficient predictive models.

The process of feature selection is not just a technical challenge; it also involves exploring a deep understanding of the data[4]. By identifying which features are most relevant to the target variable, data scientists can better understand the patterns and structures behind the data. This understanding is crucial for building models that accurately reflect real-world situations and is one of the reasons feature selection holds a central position in machine learning. In practical applications, feature selection can also help us save a significant amount of time and resources. By reducing the amount of data that needs to be processed during model training and prediction, we can iterate models more quickly and respond to market changes more rapidly. Additionally, feature selection contributes to improving the portability of models, enabling them to maintain their predictive capabilities across different datasets and environments, which is particularly important in fields that need to adapt quickly to new situations.

Feature selection is not just about reducing the number of features but also about enhancing the signal-to-noise ratio within the data. By focusing on the most informative features, we can more accurately capture the underlying patterns that drive the outcomes we're interested in. This refinement process is crucial for improving the model's robustness and generalizability across various data environments. The efficacy of feature selection is particularly evident when models are deployed in real-world scenarios where data can be noisy and inconsistent. A well-selected feature set can lead to more robust models that are less sensitive to the idiosyncrasies of a particular dataset, thus providing more reliable predictions. Moreover, feature selection plays a critical role in managing the computational complexity of machine learning models. In fields like genomics, where the volume of data is vast, feature selection helps in identifying the genetic markers that are most relevant to a particular disease or condition, which is essential for targeted treatments and diagnostics. This process also has ethical implications, as it can help in identifying biases in the data and ensuring that the models are fair and unbiased. By carefully selecting features, we can avoid reinforcing stereotypes and ensure that the models make decisions based on relevant and objective criteria. Feature selection is also a key component in the development of explainable AI, where the ability to explain the reasons behind a model's predictions is crucial. By reducing the number of features, we can simplify the model's decision-making process, making it easier to understand and trust.

3 METHODOLOGY

The Feature selection is a critical step in machine learning, involving the identification of the most valuable subset of features from a large pool of available ones for model prediction. This process can be achieved through a

variety of algorithms, mainly divided into three categories: Filter Methods, Wrapper Methods, and Embedded Methods. Filter Methods, such as correlation coefficient and chi-squared test, assess the relationship between features and the target variable based on statistical tests, independent of any learning algorithm. By quantifying the association between features, we can identify those that are most predictive of the target variable, thus enhancing the model's performance and interpretability[3]. The integration of these filter methods with other techniques, like maximum flow and minimum cost flow theory, can further optimize complex decision-making processes, such as evacuation planning. Che et al. first inrtoduced this theory in the field. Maximum flow and minimum cost flow theory are used to solve network flow problems, where the goal is to maximize the flow from a source to a sink or to minimize the cost of the flow while satisfying capacity constraints[3]. We also adopt this method in our research. Wrapper Methods, like Recursive Feature Elimination, evaluate the importance of features by actually training models, often requiring cross-validation to determine the best subset of features. Embedded Methods, such as Lasso regression, combine the feature selection process with model training, reducing the number of features through regularization techniques. These methods have their own advantages and limitations, and the choice of the right feature selection algorithm needs to be decided based on the specific problem and the characteristics of the dataset.

In machine learning projects, the choice of feature selection algorithms directly impacts the performance of the model. Filter Methods are widely used due to their simplicity and speed, especially in scenarios where a large number of features need to be preliminarily screened. However, as they do not consider the interaction between features, they may overlook some important information. Wrapper Methods, although more computationally expensive, can provide a more accurate subset of features because they are optimized directly for a specific learning algorithm. Embedded Methods offer a compromise, conducting feature selection during the model training process, making the selection process more aligned with the needs of the final model[5]. For instance, Lasso regression achieves automatic feature selection through L1 regularization, a method particularly useful when dealing with datasets that have multicollinearity. Overall, the choice of feature selection algorithms should be based on an understanding of the problem, the characteristics of the data, and the consideration of computational resources[6].

3.1 Statistical Methods

Statistical methods select features by evaluating the statistical relationship between features and the target variable. For instance, the Chi2 test is used for classification problems, selecting features by testing their independence from the target variable[7]. Pearson Correlation measures the linear relationship between features and the response variable, with values ranging from -1 to 1. Mutual Information is a non-parametric method that can assess any type of relationship and is suitable for sparse data. These methods are simple and fast but may not be suitable for all types of data, especially when relationships are non-linear.

Statistical methods are among the earliest techniques used in feature selection, evaluating the statistical relationship between features and the target variable to select features[8]. These methods include ANOVA, Chi-squared test, Pearson correlation coefficient, and mutual information. ANOVA is used to assess whether the differences between features and the target variable are significant, applicable to classification and regression problems. The Chi-squared test is another popular statistical method, especially suitable for classification problems, selecting features by testing their independence from the target variable[9]. Pearson's correlation coefficient measures the linear relationship between features and the response variable, while mutual information is a non-parametric method suitable for assessing any type of relationship, especially for sparse data. These statistical methods are simple and fast, suitable for preliminary screening of a large number of features. By combining the statistical rigor of feature selection with the optimization capabilities of flow theories, we can develop more robust and explainable models that not only perform well but also provide clear insights into the decision-making process, which is essential for regulatory compliance and ethical use in business management .

However, they may not be suitable for all types of data, especially when the relationship between features and the target variable is non-linear. In addition, these methods independently evaluate each feature and cannot consider the interaction between features, which may lead to the selection of a suboptimal feature subset[10].

3.2 Model-based Methods

Model-based methods use machine learning algorithms to assess the importance of features. Random Forest evaluates feature importance by building multiple decision trees and calculating the average reduction in impurity for each feature. Lasso regression uses L1 regularization to push less important feature coefficients towards zero, achieving feature selection[11]. These methods provide more accurate feature selection as they consider the contribution of features to the model's predictive power. However, they are generally more computationally expensive than statistical methods and depend on the performance of the selected model.

Model-based methods are a powerful tool for feature selection, relying on specific machine learning algorithms to assess the importance of features. Random Forest is a popular model-based method that builds multiple decision trees and calculates the average reduction in impurity for each feature to assess feature importance. Cheng et al. introduced a scoring mechanism that significantly enhanced the performance of the algorithm[4]. This method can provide a global importance score for features, suitable for problems that need to consider the interaction between features. Lasso

regression is another model-based method that achieves feature selection by pushing less important feature coefficients towards zero through L1 regularization. This method is particularly suitable for regression problems. Model-based methods can provide more accurate feature selection as they consider the contribution of features to the model's predictive power[12-13]. In addition, these methods may require tuning the model's hyperparameters to obtain the best feature selection results.

3.3 Recursive Feature Elimination

Recursive Feature Elimination is a wrapper method that works by recursively training a model and removing the least important features[7]. It starts with an initial set of features and repeatedly selects and removes the least important features until the desired number of features or performance criteria are met. This method allows us to perform feature selection directly for a specific model and can be used with any classifier or regressor[7]. A key advantage of RFE is that it provides feature ranking, which helps understand which features contribute the most to the model's predictive power and is essential for regulatory compliance and ethical use in business management .

Recursive Feature Elimination is a powerful wrapper method for feature selection that works by recursively training a model and removing the least important features. RFE starts with an initial set of features and then repeatedly selects and removes the least important features until the desired number of features or performance criteria are met. This method allows us to perform feature selection directly for a specific model and can be used with any classifier or regressor. A key advantage of RFE is that it provides feature ranking, which helps understand which features contribute the most to the model's predictive power. However, RFE may require high computational costs, especially when the number of features is large, as it requires training the model multiple times. Despite this, RFE remains a powerful tool that can help identify and eliminate features that contribute less to model performance. RFE can also be combined with other model selection techniques, such as cross-validation, to find the optimal number of features. The flexibility and effectiveness of this method make it a popular approach in feature selection[14].

3.4 Automated Feature Engineering

The Automated feature engineering tools such as Featuretools and Auto-Sklearn provide an automated process for feature generation and selection[15]. Featuretools focuses on time series data, automatically discovering relationships between entities and generating features. Auto-Sklearn is an automated machine learning toolkit that integrates feature selection, model selection, and hyperparameter tuning, capable of automatically selecting the best machine learning pipeline. These tools significantly improve the efficiency and scalability of machine learning projects by automating complex feature engineering tasks.

Automated feature engineering tools greatly simplify the process of feature selection and generation, helping data scientists save a significant amount of time and resources through automation[16-17]. Featuretools is an open-source library specifically designed for automated feature engineering, capable of handling complex data relationships and generating meaningful features. Featuretools uses Deep Feature Synthesis (DFS), a recursive method that explores relationships in data and creates new features[18]. This method is particularly suitable for dealing with time series data and multi-table datasets. By combining the statistical rigor of feature selection with the optimization capabilities of flow theories, we can develop more robust and explainable models that not only perform well but also provide clear insights into the decision-making process, which is essential for regulatory compliance and ethical use in business management. We duplicated the experients using the same algorithm, and it has proven great efficiency. Auto-Sklearn is an automated machine learning tool that not only includes feature selection but also integrates model selection and hyperparameter tuning. Auto-Sklearn can automatically select the best machine learning pipeline, including preprocessing, feature selection, model selection, and model tuning. The automation features of these tools make it easy for even non-experts to apply complex feature engineering techniques, improving the efficiency and scalability of machine learning projects. However, these tools may require considerable computational resources and may encounter performance bottlenecks when dealing with very large datasets. It is essential for regulatory compliance and ethical use in business management. In addition, the results of automated feature engineering tools may need further verification and adjustment by domain experts to ensure that the generated features are closely related to the actual problem.

4 CONCLUSION

In conclusion, the integration of machine learning and feature selection in business management has proven to be a transformative approach, offering significant advancements in decision-making, operational efficiency, and strategic planning. By harnessing the power of algorithms to identify the most relevant features from vast datasets, companies can streamline their processes, reduce costs, and enhance the accuracy of their predictions and forecasts. This paper has underscored the importance of feature selection in refining machine learning models, which is crucial for handling the complexity and volume of data typical in business environments.

The applications discussed have demonstrated that feature selection not only improves the performance of machine learning models but also provides insights into the underlying patterns that drive business outcomes. From customer segmentation to supply chain optimization, the strategic use of feature selection has enabled businesses to make more informed decisions, leading to competitive

advantages in the market.

Looking forward, the continued evolution of machine learning techniques, coupled with the growing sophistication of feature selection methods, promises to unlock further potential in business management. As businesses increasingly rely on data-driven strategies, the ability to extract meaningful information from complex data sets will be paramount. Therefore, investing in the development and deployment of advanced feature selection techniques will be a key differentiator for organizations seeking to stay ahead in the rapidly evolving landscape of business management.

In summary, the intersection of machine learning and feature selection presents a compelling opportunity for businesses to leverage data analytics for enhanced decision-making and operational excellence. As this field continues to mature, the implications for business management are profound, suggesting a future where data-driven insights become the cornerstone of strategic business practices.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

CHEN, Siyao

University of Connecticut, USA.

DAIYA, Abhit

University of Connecticut, USA.

DEB, Rubhrat

Wildlife Institute of India, India.

REFERENCES

[1] Che, C., & Tian, J. (2024). Game Theory: Concepts, Applications, and Insights from Operations Research. Journal of Computer Technology and Applied Mathematics, 1(4), 53-59.

[2] Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), 157-170.

[3] Che, C., & Tian, J. (2024). Maximum flow and minimum cost flow theory to solve the evacuation planning. Advances in Engineering Innovation, 12, 60-64.

[4] Cheng, X. (2024). A Comprehensive Study of Feature Selection Techniques in Machine Learning Models.

[5] Taha, A., Cosgrave, B., & Mckeever, S. (2022). Using feature selection with machine learning for generation of insurance insights. Applied Sciences, 12(6), 3209.

[6] Chen, L. H., & Hsiao, H. D. (2008). Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study. Expert systems with applications, 35(3), 1145-1155.

[7] Cheng, X. (2024). Machine Learning-Driven Fraud Detection: Management, Compliance, and Integration. Academic Journal of Sociology and Management, 2(6), 8-13.

[8] Di Mauro, M., Galatro, G., Fortino, G., & Liotta, A. (2021). Supervised feature selection techniques in network intrusion detection: A critical review. Engineering Applications of Artificial Intelligence, 101, 104216

[9] Che, C., & Tian, J. (2024). Methods comparison for neural

network-based structural damage recognition and classification. Advances in Operation Research and Production Management, 3, 20-26.

[10] Zhao, Z., Anand, R., & Wang, M. (2019, October). Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In 2019 IEEE international conference on data science and advanced analytics (DSAA) (pp. 442-452). IEEE.

[11] Che, C., & Tian, J. (2024). Understanding the Interrelation Between Temperature and Meteorological Factors: A Case Study of Szeged Using Machine Learning Techniques. Journal of Computer Technology and Applied Mathematics, 1(4), 47-52.

[12] Tian, J., & Che, C. (2024). Automated Machine Learning: A Survey of Tools and Techniques. Journal of Industrial Engineering and Applied Science, 2(6), 71-76.

[13] Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. Knowledge and Information Systems, 53, 551-577.

[14] Che, C., & Tian, J. (2024). Leveraging AI in Traffic Engineering to Enhance Bicycle Mobility in Urban Areas. Journal of Industrial Engineering and Applied Science, 2(6), 10-15.

[15] Cheng, X. (2024). Investigations into the Evolution of Generative AI. Journal of Computer Technology and Applied Mathematics, 1(4), 117-122.

[16] Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. Information & management, 39(3), 211-225.

[17] Cheng, X., & Che, C. (2024). Optimizing Urban Road Networks for Resilience Using Genetic Algorithms. Academic Journal of Sociology and Management, 2(6), 1-7.

[18] Cheng, X., & Che, C. (2024). Interpretable Machine Learning: Explainability in Algorithm Design. Journal of Industrial Engineering and Applied Science, 2(6), 65-70.